JOURNAL OF ENGINEERING RESEARCH & SCIENCES



Special Issue

Multidisciplinary Sciences and Advanced Technology

June 2023

www.jenrs.com
ISSN: 2831-4085

JENRS

EDITORIAL BOARD
(Special Issue)

Guest Editor

Prof. Paul Andrew
Department of Electrical Engineering, Universidade De São Paulo, Brazil

Editorial

The growing complexity of contemporary global challenges has made it increasingly clear that meaningful progress can no longer be achieved within the confines of isolated disciplines. It is in this context that we are pleased to present this special issue of the *Journal of Engineering Research and Sciences*, devoted to *Multidisciplinary Sciences and Advanced Technology*. This issue brings together a diverse body of research that reflects the evolving nature of scientific inquiry one that is collaborative, integrative, and driven by technological advancement.

Multidisciplinary research has become a defining feature of modern innovation. By fostering collaboration across engineering, natural sciences, computational domains, and applied technologies, researchers are able to develop holistic solutions to problems that are inherently complex and interconnected. The contributions included in this special issue demonstrate how the convergence of diverse fields leads to new perspectives, innovative methodologies, and impactful outcomes. Whether addressing challenges in energy systems, healthcare technologies, smart infrastructure, or environmental sustainability, the studies presented here underscore the importance of breaking traditional academic boundaries.

A central focus of this issue is the role of advanced technology as a catalyst for multidisciplinary integration. Emerging technologies such as artificial intelligence, machine learning, the Internet of Things (IoT), and nanotechnology are reshaping the research landscape. These tools enable seamless interaction between disciplines, allowing for the development of intelligent systems that are adaptive, efficient, and scalable. The research featured in this issue highlights how these technologies are being leveraged to enhance system performance, optimize resource utilization, and support data-driven decision-making across a wide range of applications.

Equally significant is the emphasis on sustainability and responsible innovation. As the world faces pressing environmental and societal challenges, there is a growing need for solutions that balance technological progress with ecological and social considerations. Several contributions in this issue explore sustainable engineering practices, renewable energy technologies, and environmentally conscious design approaches. By integrating insights from multiple disciplines, these studies offer pathways toward achieving long-term sustainability and resilience.

The importance of computational tools and data-centric approaches is also a recurring theme throughout this issue. Advanced modeling, simulation, and data analytics have become essential in understanding complex systems and predicting their behavior. These methodologies enable researchers to test hypotheses, optimize designs, and accelerate the translation of theoretical concepts into practical solutions. The integration of computational techniques with experimental and theoretical research exemplifies the transformative potential of multidisciplinary approaches.

The editorial team extends its sincere appreciation to all authors for their valuable contributions and to the reviewers for their diligent and rigorous evaluations. Their expertise and commitment have ensured the quality and integrity of this special issue, making it a meaningful contribution to the broader research community.

As we present this collection of scholarly work, we hope it will serve as a platform for continued dialogue, collaboration, and innovation. The future of science and technology depends on our ability to integrate knowledge across disciplines and to harness advanced technologies in addressing global challenges. We encourage readers to engage with the insights presented in this

issue and to contribute to the ongoing advancement of multidisciplinary sciences and advanced technology.

Guest Editor

Prof. Paul Andrew

JOURNAL OF ENGINEERING RESEARCH AND SCIENCES

Special Issue **June 2024**

CONTENTS

01 *SimulatorBridger: System for Monitoring Energy Efficiency of Electric Vehicles in Real-World Traffic Simulations*
by Reham Almutairi, Giacomo Berg and Graham Morgan

02 *A Thorough Examination of the Importance of Machine Learning and Deep Learning Methodologies in the Realm of Cybersecurity: An Exhaustive Analysis*
by Ramsha Khalid and Muhammad Naqi Raza

03 *A Computational Approach for Recognizing Text in Digital and Natural Frames*
by Mithun Dutta, Dhonita Tripura and Jugal Krishna Das

04 *Dynamic and Partial Grading of SQL Queries*
by Benard Wanjiru, Patrick van Bommel and Djoerd Hiemstra

05 *MCNN+: Gemstone Image Classification Algorithm with Deep Multi-feature Fusion CNNs*
by Haoyuan Huang and Rongcheng Cui

06 *The Dual Impact of AI in Clinical Trials: Perspective*
by Muhammad Yaqub and Lan He

07 *Fingerprint Bio-metric: Confronting Challenges, Embracing Evolution, and Extending Utility – A Review*
by Diptadip Maiti, Madhuchhanda Basak and Debashis Das

08 *HivePool: An Exploratory Visualization System for Honey Beehive Data*
by Tinghao Feng, Sophie Columbia, Christopher Campell and Rahman Tashakkori

09 *An Integrated Approach to Manage Imbalanced Datasets using PCA with Neural Networks*
by Swarup Kumar Mondal and Anindya Sen

10 *Conceptual Business Model Framework for AI-based Private 5G-IoT Networks*
by Laurence Banda

11 *Navigating the Autonomous Era: A Detailed Survey of Driverless Cars*
by Vaibhavi Tiwari

12 *Evaluation of Equivalent Aacceleration Factors of Repairable Systems in a Fleet: a Process-Average-Based Approach*
by Renyan Jiang, Kunpeng Zhang, Xia Xu and Yu Cao

13 *Enhancing Mental Health Support in Engineering Education with Machine Learning and Eye-Tracking*
by Yuexin Liu, Amir Tofighi Zavareh and Ben Zoghi

14 *Secure Anonymous Acknowledgments in a Delay-Tolerant Network*
by Edoardo Biagioni

15 *Fuzzy-Based Approach for Classifying Road Traffic Conditions: A Case Study on the Padua-Venice Motorway*
by Gizem Erdinc, Chiara Colombaroni and Gaetano Fusco

16 *A Comparative Analysis of Interior Gateway Protocols in Large-Scale Enterprise Topologies*
by Saleh Hussein Al-Awami, Emad Awadh Ben Srity and Ali Tahir Abu Raas

17 *Biclustering Results Visualization of Gene Expression Data: A Review*
by Haithem Aouabed, Mourad Elloumi and Fahad Algarni

18 *Advanced Cloud-Based Solutions for Peripheral Artery Disease: Diagnosis, Analysis, and Visualization*
by Mohammed A. AboArab, Vassiliki T. Potsika and Dimitrios I. Fotiadis

19 *Advanced Digital Twin of a Industrial Robotic System for Measuring Pipe Wall Thickness in Nuclear Power Plants*
by Rogério Adas Pereira Vitalli and João Manoel Losada Moreira

20 *Developing E-Methane Value Chain and Proper Greenhouse Gas (GHG) Accounting Rules Incentivizing Recycled Carbon Fuels*
by Ryota Kuzuki, Mitsuhiro Kohara, Noboru Kizuki, Satoshi Yoshida, Yuta Nezasa and Yuki Tsuji

21 *How to Analyze the Introduction Strategies for Radically New Technological Innovations? The Case of Autonomous Shipping*
by Mariah Kurtinaitis Joukes, Roland Ortt, Mark de Bruijne and Linda M. Kamp

22 *Enhancing Python Code Embeddings: Fusion of Code2vec with Large Language Models*
by Long H. Ngo and Jonathan Rivalan

Received: 22 May, 2024, Revised: 19 June, 2024, Accepted: 24 June, 2024, Online: 30 June, 2024
DOI: https://dx.doi.org/10.55708/js0306004

SimulatorBridger: System for Monitoring Energy Efficiency of Electric Vehicles in Real-World Traffic Simulations

Reham Almutairi[*,1,2], **Giacomo Bergami**[2], **Graham Morgan**[2]

[1]College of Computer Science and Engineering, University of Hafr Albatin, Saudi Arabia
[2] School of Computing, Newcastle University, United Kingdom
[*]Corresponding author: Reham Almutairi, reham.mutlaq@hotmail.com

ABSTRACT: The increasing popularity and attention in Vehicular Ad-hoc Networks (VANETs) have prompted researchers to develop accurate and realistic simulation tools. Realistic simulation for VANETs is challenging due to the high mobility of vehicles and the need to integrate various communication modalities such as Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) interactions. Existing simulators lack the capability to simulate VANET environments based on IoT infrastructure. In this work, we propose SimulatorBridger, a novel simulator that bridges IoTSim-OsmosisRES with SUMO, a traffic simulator, to simulate VANET environments with integrated IoT infrastructure. Our study focuses on analyzing the generated dataflows from V2I and V2V interactions and their impact on vehicle energy efficiency. Even though On-Board Units (OBUs) appear to have insignificant energy demands compared to other vehicle energy consumptions such as electric motors or auxiliary systems (HVAC, lights, comfort facilities), we found a near-perfect correlation between the intensity of communication dataflows and the battery consumption. This correlation indicates that increased communication activity can contribute to an increase in overall energy consumption. Furthermore, we propose future research directions, including traffic rerouting based on battery consumption optimization, which can be efficiently tested using our simulation platform. By including communication energy costs in the design of energy-efficient vehicular networks, these insights contribute to a deeper understanding of energy management in VANETs.

KEYWORDS Traffic Simultors, VANET, IoT

1. Introduction

The advancement of technology leads us towards a new area of communication connecting two or more vehicles to exchange data within an IoT ecosystem. This drives us through a new VANET (Vehicular Adhoc Network) framework making the conventional transport system safer, full of infotainment, convenient, and smart. In VANET research, researchers aim to develop a simulator platform that provides *quick* and *cost-efficient* transmission of data for passenger safety and comfort. As described earlier, VANET simulators consist of two components: Traffic Simulators (software environments that generate vehicle movements in trace files) and Network Simulators (used to test the performance of networking protocols). The network simulators are used to build communication topologies, evaluate network protocols, and exchange routing information between the nodes after importing the traces of mobility models generated by traffic simulators. These mobility models, which are the depiction of real-world scenarios, are called *traces* and contain the simulated infrastructure and event information such as vehicle speed, type, origin, destination, arrival time, arrival rate, maximum density, number of lanes, speed limits, capacity, intersection type, queuing, service distribution, service rate, traffic signs, location, etc. Green and orange paths in Figure 1a provide a minimalistic view of these traces where only the vehicle's geographical position is considered. The traces generated with the mobility generator, microscopic or macroscopic [1, 2], are imported into the network simulator and generate a vehicular program: boxes surrounding paths in Figure 1b provide a depiction of the vehicular programs associated with simulated IoT agents.

1.1. Objective and motivation

We have used IoTSim-OsmosisRES [3] as the IoT simulator for our proposed coupling platform, as it is the only IoT simulator considering energy management, a variety of power sources, and network infrastructure, while others can neither simulate floating weather conditions, nor renewable energy sources. However, the current implementation of the IoT simulator, *IoTSim-OsmosisRES* assumes IoT devices to be in fixed locations. So, IoT device mobility is not supported. Hence, we propose a new simulator SimulatorBridger that couples a state-of-the-art IoT simulator (IoTSim-OsmosisRES [3]) with a traffic simulator (SUMO [4]) to simulate VANET environments. This proposed simulator is designed to meet these goals: coupling the IoT simulator IoTSim-OsmosisRES with the traffic simulator, SUMOU, enhancing IoTSim-OsmosisRES capabilities to be

able to simulate moving IoT devices and to support battery information associated to the IoT devices (vehicles).

Our proposed simulator allows a smooth integration of mobility, IoT devices, heterogeneity, and battery management in highly heterogeneous and dynamic environment. Still, we demonstrate that our architecture has sufficient flexibility to bridge any possible traffic simulator using a plug-and-play approach. We provide all the necessary functionality for the accurately couple simulators that can meet the complexity of traffic environments by using the most recent IoT simulator and the first simulation framework that offers unified modelling and simulation of complex IoT applications over heterogeneous edge-cloud environments. Preliminary results (§5) remark the partial correctness of our simulator, as the vehicular battery consumption distribution follows the same trend as packet distribution. By limiting the communications to exchanging one single packet between the cloud and the IoT node, we also show that the number of communications follows a similar distribution to the former. We finally show that increasing the number of packets travelling in the network increases the overall time required by the network to process these, as bottlenecks might occur in the network simulator infrastructure. Future work will focus on testing different communication policies aiming to decrease the packet processing overhead within the network. These preliminary fidelity results remark the adequacy of using the simulator as a *digital twin* for a realistic urban mobility scenario, thus allowing any policy-maker to test disparate network configuration and traffic distributions to minimise network communication overload.

1.2. Use Case Scenario

The recent Car-as-a-Service paradigm [5] has remarked the possibility of using the cars' IoT sensors not only for the personal purposes of travellers but, with the driver's agreement, for the benefit of the entire community. As most modern cars are equipped with navigation systems and video cameras, cars can collect videos and images of their surroundings and geo-tag such information. This might be extremely useful for any town hall highly committed to road surface remaking for car safety so that they can spot tarmac conditions before they degenerate into potholes [6]. To do so, the data needs to be collected from cars and then streamed towards the main data centre, Figure 1, where an AI model is going to detect the seriousness of the situation [7]. Road-Side Units (RSU) might first collect the data from the cars through 5G antennae supporting massive low-latency communication and stream it towards the primary data centre thanks to the core networks. Communication between 5G antennae is supported by optical fibres realising low-latency communications. As a result of the data collection process, a massive amount of data will stream every second from the cars towards the data centres, which will severely increase during rush hours when road congestions might happen with higher probability [8]. Figure 2 represents a portion of the simulated traffic from Hamburg [1]. Due to the massive volume of data generated by the high number of vehicles during rush hours, any road congestion will constitute a communication bottleneck degrading the overall communication performance [9]. To achieve success in these situations, the traffic load needs to be balanced [10] to reduce communication delays. Given that the traffic flow might be redirected at run-time, this motivates the representation of each RSU as a specific Edge [3]. The exploitation of Osmotic Computation [11] eases communication flow management, thus dynamically establishing new streams. This also motivates the definition of a simulator bridging traffic simulation to network communication systems.

2. Design of the Simulator

Our framework twins traffic simulation and IoT simulation. Thus, it incorporates well-known models for road traffic micro-simulation with a comprehensive selection of models of network protocols. By running the vehicular generator, we collect all the vehicular geographical information and the position of the Edge node (§2.1). This is required to set up the network topologies associated with each Edge node and determine the vehicular program to be injected into the IoT devices in the IoT simulator (§2.2). At each simulation time, each IoT device will query a centre to determine whether the IoT device shall to establish communication with the cloud through an edge device or not (§2.3). Upon confirmation, the IoT device directly contacts the MEL router associated with the edge node of choice, establishing communication between the IoT device and the cloud through one of its hosts (§2.4).

2.1. Vehicular Data Collection

First, the simulator collects the mobility information generated by the traffic generator. This simulator might be configured by setting the begin τ_b and the end τ_e simulation time, as well as a temporal granularity δ defining the sampling rate of the vehicular information. This allows do determine the time at the i-th simulation step for each vehicle v as the following relation:

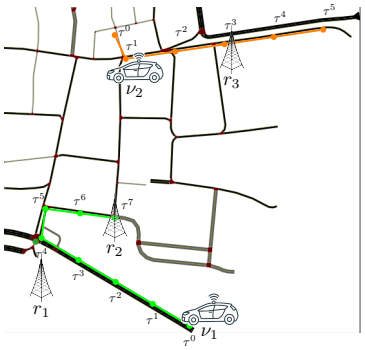
$$\tau_v^i \begin{cases} \tau_v^{i-1} \ \delta & i > 0 \\ \tau_b & \text{oth.} \end{cases}$$

At the end of the traffic generation, the VANET simulator returns a list of pairs associating each simulation time τ_v^i to a geographical location ϖ_v^i for each vehicle v. Figure 1a provides a graphical depiction of such traces, highlighted in either green or orange. From the network topology, we might also infer the geographical position of each RSU of interest r alongside its position ϖ_r and its communication radius ρ_r. This operation is performed by a simulator wrapper, which provides a standard interface for generating the data of interest in a uniform representation independently from the specific traffic simulator of choice.

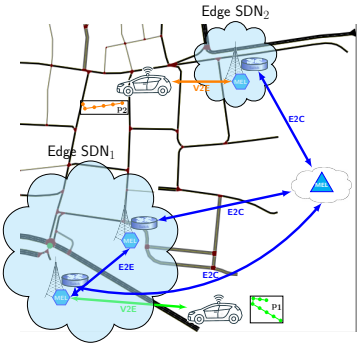
2.2. Software Defined Network Configuration

The next step assumes that each vehicle v is associated with just one single IoT device potentially communicating with the cloud infrastructure. We collect the sequence of pairs

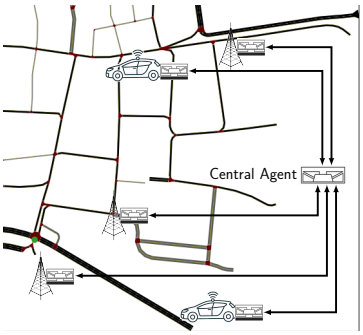
[1] https://github.com/DLR-TS/sumo-scenarios/tree/main/TAVF-Hamburg



(a) Output of the SUMO simulator, associating each vehicle v to a trace (paths in green or orange) determining their location ϖ_v^i at each simulation frame τ^i. RSU r are associated with fixed geographical coordinates ϖ_r.

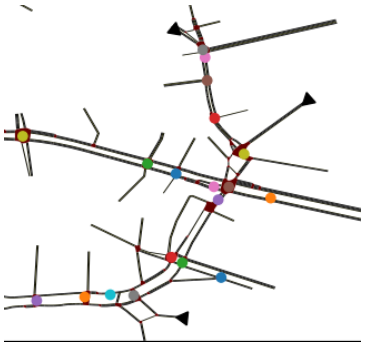


(b) Osmotic Applications (MEL representation) remarking the Vehicle2Edge (V2E), Edge2Edge (E2E), and the Edge2Cloud (E2C) communication channels. Boxes denoted as p_1 and p_2 associated to each vehicle indicate the "simulation programs" retaining the mobility information.

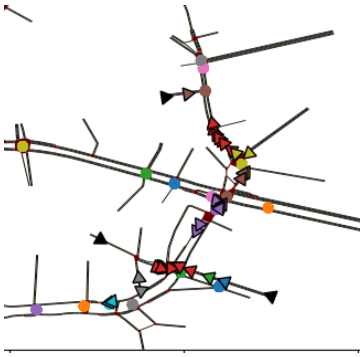


(c) Supporting the Osmotic Agent configuration from IoT-Osmosis-RES with centralized training and decentralized execution for dynamically updating the communication strategies of the IoT devices. This enables us to support dynamic load balancing in the future.

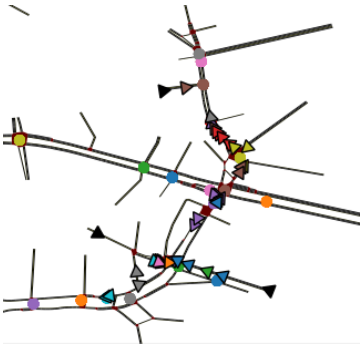
Figure 1: Osmotic Computing in car traffic scenario in Newcastle Upon Tyne.



(a) Some road lanes (black) and some RSU (colored circles). Black triangles are vehicles not associated with an RSU.



(b) Associating each semaphore with the nearest cars within the radius of θ : cars and RSU with the same colour communicate.



(c) After exploiting a load balancing algorithm, the cars are distributed to other semaphores in the neighbourhood.

Figure 2: A subset of the Sumo TAV Hamburg Dataset for mobility

$\left[\left(\tau_v^0, \varpi_v^0\right), \ldots, (\tau_v^n, \varpi_v^n)\right]$, where τ^n τ_e, as a vehicular program v that will be injected within the IoTOsmosis-RES simulator.

This will allow the IoT device to update its position at each instant of the SUMO simulation time in the network simulator. Fig. 1b represents such vehicular programs as minimizations of the traces represented in Fig. 1a.

Contextually, we define Software-Defined Wide Area Networks for the RSU nodes. To do so, we first define an undirected graph G V, E, where the vertices V list the RSU nodes r obtained from the VANET simulator, and the edges establish a communication channel between two distinct RSU r and r' if and only if they can both communicate between each other:

$$\forall r, r' \in V. \quad (r, r') \in E \leftrightarrow r \neq r' \wedge \|\varpi_r - \varpi_{r'}\| \leq \min(\rho_r, \rho_{r'})$$

For each strongly connected component of such graph, we establish a distinct Edge SDN with an associated Datacenter and Software-Defined Network Controller, where each RSU is described as an Edge device associated with at least one host. Each light blue cloud in Fig. 1b represents a distinct Edge SDN, for which only the Edge nodes are remarked. We also set up a Cloud network towards which each IoT device will communicate to send the sensed data, as well as an SDWAN network bridging each Edge SDN to a Cloud SDN. Both these structures are engulfed in the white cloud in Fig. 1b. Last, we completely set up the edge connectivity between these networks by setting up Edge2Edge and Edge2Cloud communication links. VehicleToEdge links will be established at simulation time.

2.3. Osmotic agents module

The osmotic agents' module is exploited for simulating network discovery procedures and determining each IoT device which should be the best edge node for establishing the communication. This cannot be necessarily determined at the local level, as each IoT device has only the information of which edges it can communicate with, but it doesn't know which is going to be the best neighbouring agent towards which establish the first-mile communication towards for reaching the cloud. An IoT device acts as an reinforcement agent, which sends the information about the observed environment to the central agent. In particular, each IoT device sends its geographical position, while the geographical position of each edge device is known a priori. We implemented two specific types of central agents:

- NearestCentralAgent: for each querying IoT device, the central agent will always respond with the nearest Edge device to the IoT that is within mutual signal coverage. This strategy does not require any alteration to the SDN Routing Policy of each Edge Network, which might still exploit the SHORTESTPATHMAXIMUMBANDWIDTH discussed in [12].

- FlowDataCenterAgent: after associating each IoT device to a possible edge device, the agent runs a multi-source and multi-target minimum-cost flow problem for establishing the best strategy to minimize the channel communication. This algorithm will then return for each IoT device the best edge device. This also requires updating the SDN Routing Policy MaximumFlowRoutingPolicy associated with each network via its SDN Controller, thus returning the paths calculated by the Central Agent.

Upon reception of a non-empty edge device, the IoT will generate a novel Osmotic Application and start communication towards the cloud through the elected edge node. Figure 1c illustrates the bidirectional communication flow between IoT or SDN Rounters and the Central Agent.

2.4. Dynamic Adaptive Routing

In osmotic computing, each IoT device directly interacts with MicroELements (MEL) software components that might be instantiated anywhere in a given network associated to Edges [3]. To faithfully represent real communication scenarios, an IoT device must directly establish communication with a precise edge node [13]: the Adaptive Routing in [3] was further extended to directly select the MEL associated with a given edge node instead of picking one in a round-robin fashion. As IoT-OsmosisRES further associates a MEL to one or more hosts, the routing algorithm picks one of the hosts running the MEL as its first-mile communication point. Contextually, the host establishes the communication channel with the cloud network indicated by the IoT device, thus starting the communication.

3. Implementation in Java

The proposed simulator, SimulatorBridger, is programmed using the Java programming language, and it is available in Github [2]. The NetworkTopology.java class in the SimulatorBridger-IOTSimOsmosisRES module is crucial for setting up the network layer in the simulation. It generates a topological network, which is used to simulate latency in network traffic. The IoTDevice.java class represents IoT devices within the IoTSimOsmosisRES framework. The CloudletScheduler class is responsible for scheduling tasks in the cloud. And EdgeDeviceManager class is managing edge devices.

The implementation of the SimulatorBridger platform contains several modules and a large number of Java source files. Here's an overview of its structure:

SimulatorBridger-core component contains shared dependencies and core functionalities used across the project. It serves as the foundational codebase that other modules in the project rely on.

SimulatorBridger-traffic-information-collector involves running the traffic simulator and collecting data from it. It processes the output of the simulation, identifying IoT nodes and Edge nodes. The Edge nodes are used for interactions by the IoT nodes.

SimulatorBridger-central-agent-planner provides a theoretical omniscient algorithm capable of scheduling time as required. Depending on the network simulator chosen, it also generate potential network connectivity information based on the IoT and Edge information provided by the simulation.

SumoOsmosisBridger is an example that bridges all the simulations together with Dynamic IoTSimOsmosisRES simulator. It illustrates how the components can be integrated in a seamless manner.

Each of these components plays a specific role in integrating and processing traffic data within the IoTSimOsmosisRES environment. The process starts with data collection and processing, followed by planning and scheduling using the central agent planner, and finally, the integration of all these elements through the SumoOsmosisBridger to work with Dynamic IoTSimOsmosisRES. This integration allows for a comprehensive simulation environment that can handle complex IoT and traffic data scenarios.

4. Dataset and Experiment Setup

This section covers running the Bologna dataset [3], through our proposed simulator, where vehicles have an embedded IoT device and RSU are Edge nodes associated to MELs. Fig. 3 shows the SUMO network of this dataset. The dataset has 16 RSUs and 239 vehicles. The Bologna scenario includes the area around the football stadium and was set up to simulate the mobility of big events such as football matches or concerts.

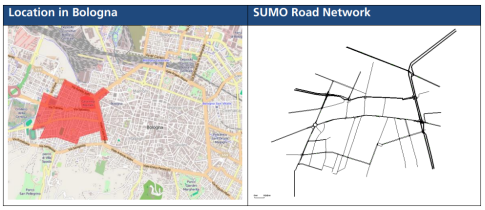


Figure 3: Bologna Dataset [14]

Our experiments orchestrate a traffic simulator (SUMO) with a network simulator (IoTOsmosis-RES) with each simulation second in the former corresponding to a second in the latter. We arbitrarily set the IoT battery consumption rate to 1.02% when a vehicle communicates with a MEL. We assumed embedded IoT devices cannot be re-charged to better analyse the simulator's correctness by correlating the number of packets sent by an IoT device with its battery consumption. When a vehicle was near an RSU in the traffic simulator, we scheduled a new communication between the IoT device embedded in the vehicle and the Cloud via the Edge node representing the RSU in the network simulator. This schedule is selected according to the granularity δ of the traffic simulator: in our scenario, this is set to start every 1 second. In this time frame, the IoT device in the vehicle sends one single packet. The simulation time for these plots begins at the start of the traffic simulator and ends when the last MEL has successfully sent all the packets to the cloud.

[2] https://github.com/jackbergus/SimulatorBridger/releases/tag/v0.1
[3] https://github.com/DLR-TS/sumo-scenarios/tree/main/bologna/acosta

Our simulator assumes each RSU is an Edge device containing multiple MELs. As per the previous discussion, the network simulator was extended to establish direct communication between IoT and Edge devices by resolving an available MEL associated with the Edge device. A round-robin policy selects an available MEL in the Edge node to ease each MEL's workload. For this experiment, the maximum simulation time of IoTOsmosis-RES was 3898.4s, corresponding to 100% of the simulation, starting at the same time, the maximum time for the SUMO simulator was 186.7s (τ_e), with vehicles no longer able to enter the urban environment after around 55% of the overall traffic simulator time, or 100s. From this, as the number of vehicles in the simulation decreases, the number of starting communications will also decrease, resulting in a normal distribution for both the number of communicating vehicles and the number of packets sent. The traffic simulation takes less than 5% of the total simulation time. This is due to packet network delays which delayed communication patterns even though the IoT devices sent no new packets. We consider all communications between an IoT vehicle and a MEL.

5. Analysis and Results

In this section, we analyze and discuss the key findings from the Bologna dataset experiment conducted on the SimulatorBridger platform. The results of our simulation are meticulously examined, focusing on critical performance metrics and the effectiveness of various strategies implemented throughout the process. This analysis not only highlights the successes and challenges encountered but also provides valuable insights into the practical implications and potential improvements for future iterations.

Fig 4 shows that each of the 163 embedded IoT devices (in distinct colours) shows an almost perfect correlation between the number of packets sent (solid line) and its battery consumption (dashed). The Spearman correlation between those for each vehicle is almost 1 , with a p-Value of 0 , thus indicating a very strong correlation between these two. From Fig 5, for most of the RSUs the number of connected to each has an approximate normal-like distribution, with most plateauing after the 100s mark, which is at about the 55% mark in VANET simulation time mentioned earlier. A maximum number of 39 IoT devices were connected to a single MEL and its Edge node at any time. This figure also explicitly shows vehicles starting at an RSU, as Edge#12 has 15 connected vehicles from the very beginning of the simulation. This figure also shows the simulation time used for these plots, the simulation time starts at time 0 seconds and then stops around 190 seconds once all the RSUs have no more connected vehicles. Next, we tested if there was a strong correlation between the number of vehicles in the simulation within a given time interval, with the overall battery consumed within that same time interval, as common-sense suggests that an increase of the number of the vehicles should match an increase of communications between a IoT and Edge devices, thus reflecting in an increase of the overall battery consumption. We found no strong correlation between these two aspects as, even though more vehicles lead to an increase of communications, this does not necessarily entail that either more communications should take place (e.g., vehicles might not be in a region covered by RSUs) or even that all vehicles have the same number of communications within the same time interval.

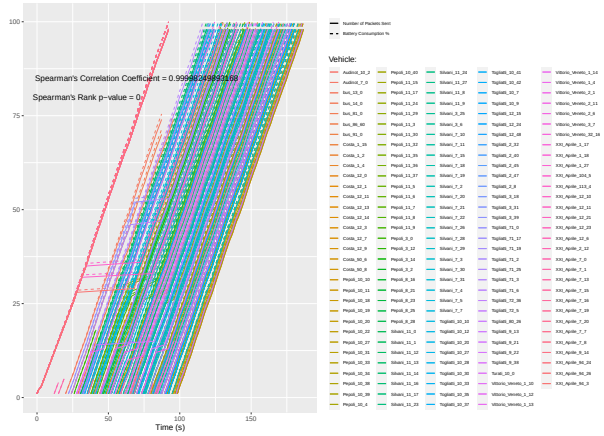


Figure 4: Correlation between battery consumption and number of packets being sent per vehicle within the simulation.

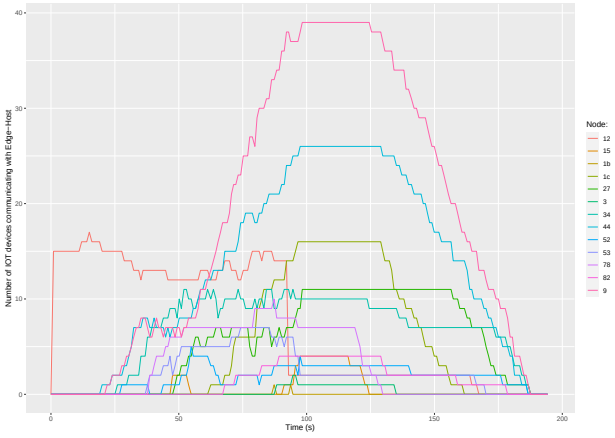


Figure 5: Number of IoT devices communicating with an Edge per simulation time.

Following this, any potential correlation between packets sent from vehicles and the overall battery consumption was investigated. Fig 6 shows the probability density of the overall battery consumed by all vehicles in the simulation (within each 5% time interval of simulation time), along with probability density of the number of packets sent by all vehicles (always within each 5% time interval of simulation time). From Fig 6 both sets of data not only closely follow a normal distribution, but also are closely correlated with each other, as both distributions and both normal plots completely overlap one another. These two variables correlating with each other was expected, as in the current simulator setup, the packets being sent are responsible for the battery consumption, and the fact that we found these two variables do in fact correlate indicates that the simulation is behaving as expected. Fig 6 shows a spike occurring at the 35 – 45% for both the overall consumption and the number of packets sent. This spike can be explained in terms of Fig 7 showing the distribution of the communications starting within each

traffic and VANET simulation time interval: this spike is also present for the number of starting communications within this same time interval. Given the current simulator settings of a vehicle sending a single packet as a result of communication with an RSU, this likefor-like behaviour of communications and packets sent is not only expected but demonstrates the simulator is functioning correctly. The reason more packets were sent and therefore why more battery was consumed in this time interval is due to more communications occurring within this time interval. Fig 7 also validates our preliminary hypothesis observing that the number of starting communications should drop after a 55% elapse of the Sumo simulation time.

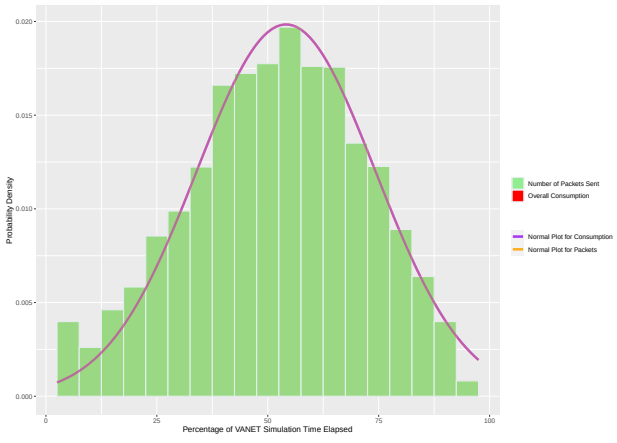


Figure 6: The distribution of packets sent in the network follows the same trend as the overall IoT battery consumption.

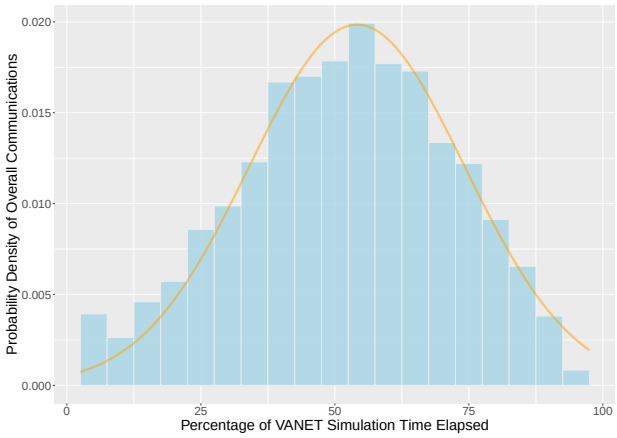


Figure 7: The distribution of the number of communications starting within each time interval.

We might also observe another spike at the beginning of the simulation where the influx of traffic starts entering the simulation while approaching RSUs, thus justifying the increase of the overall number of communications. The explanation for the initial drop-off between the first two intervals, 0 – 5% and 510%, can be explained with the first interval being artificially high as a node is placed very close to where vehicles initially join when the simulation starts. This behaviour can also be seen on Fig 5 as Edge#12 having 15 connected vehicles just 1 second into the simulation.

From Fig 8 vehicles at the start of the simulation enter from the most southwestern entrance, which, when factoring in the 100 m radius of communication of the nodes, explains how vehicles are able to connect to Edge#12 (i.e. the light blue Node 12 in the bottom left of the plot). After 190 seconds in the simulation, the Sumo simulation stops, and therefore we observed no vehicles connected to any of the RSUs. In fact, the IoTOsmosis-RES still has to wait for all the Edge nodes to send their packets to the cloud while receiving an acknowledgement for this. The discrepancy between the end of the traffic simulator and the Network simulator also matching the end of the VANET simulator has been cut off for the sake of legibility. This processing of all the packets by the network infrastructure therefore takes over 3700 seconds: we refer to this as *shutdown time*. Fig 9 shows how this shutdown time is affected by the number of total communications in the simulator: the more time steps in the simulation, or the longer vehicles are allowed to navigate in the traffic simulator, the more the communications between vehicles and RSUs take place. This matches the intuition that more time for potential communications entails an increase in the number of communications. The more relevant result from this figure with respect to the shutdown, however, shows that more communications lead to a longer and longer shutdown time.

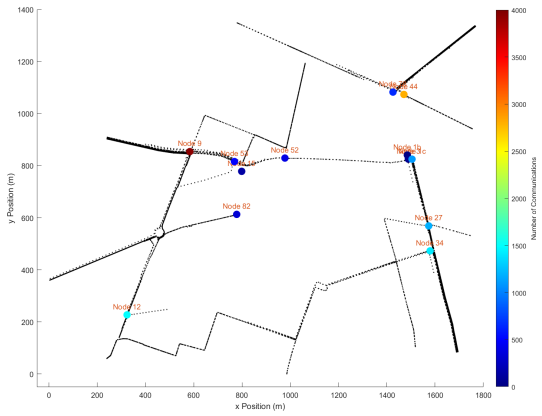


Figure 8: Showing Edge displacements as RSU in the Bologna Dataset: their colour represents the intensity of the undergoing communication. Black lines show the trajectory of the vehicles carrying embedded IoT devices.

Hence, the effectiveness of SimulatorBridger was validated through both simulation and experimental tests using the Bologna dataset. The simulation results demonstrated a near-perfect correlation between packet transmission and battery consumption, with a Spearman correlation coefficient close to 1 and a p-value of 0, indicating strong statistical significance. Experimental tests mirrored these findings, showing similar trends and validating the simulator's accuracy in modeling real-world communication patterns and their impact on energy usage. Minor discrepancies between the simulation and experimental data were observed, primarily due to real-world variances such as sensor inaccuracies and environmental factors. Statistical tests confirmed that both datasets followed similar distributions, with most errors within acceptable limits. Overall, the strong alignment between simulation and experimental re-

sults underscores the reliability of SimulatorBridger. These findings demonstrate the robustness of SimulatorBridger and its value as a tool for studying and optimizing vehicular networks.

Finally, the results offer significant improvements over previous simulation benchmarks. The correlation analysis between packet transmission and battery consumption reveals a Spearman correlation coefficient close to 1, indicating a strong relationship between communication activity and energy usage. Enhanced traffic flow and load balancing capabilities are achieved through dynamic management of communication loads among RSUs, resulting in smoother traffic flow and reduced bottlenecks. The scalability of the simulator is demonstrated by its ability to manage extensive simulations, addressing the scalability issues faced by earlier systems. Additionally, the detailed energy management features, considering various power sources and their consumption rates, set SimulatorBridger apart from traditional models that often overlooked these aspects. These findings affirm the simulator's effectiveness in providing a robust platform for studying and optimizing vehicular network performance and energy efficiency.

6. Conclusion and future works

Due to the high mobility of vehicles in VANETs, realistic simulation is a challenging task. This paper proposes a novel simulator, SimulatorBridger, by bridging the IoT network simulation with the traffic simulation. The efficacy of SimulatorBridger is validated using a case study for urban roads in Bologna city. Results show the various capabilities of SimulatorBridger in terms of vehicular network lifetime, vehicle battery, and energy consumption. According to our results, SimulatorBridger is also scalable in terms of vehicle count and simulation time. Notably, SimulatorBridger simulates VANET based on IoT infrastructure, a capability not present in existing VANET simulators, providing a more comprehensive and integrated approach to vehicular network simulation. Furthermore, our framework can be easily extended to support a load balancing scheme between traffic lights, thus minimising load imbalances in the communication network while improving energy management and implementation time. A cooperative approach for load balancing among the network might be used when the traffic light receives more requests than the maximum number, resulting in high traffic or load. Also, vehicular network lifetime is increased by reducing energy consumption which is necessary to balance energy in traffic lights. In addition, the simulator we propose combines IoTSim-OsmosisRES with any potential traffic simulator. Then for future work, a different traffic simulator can be used in conjunction with IoTSim-OsmosisRES. Fig. 9 suggests that the network infrastructure is severely affected by the communication delays by the way it processes and transmits the packets received from the IoT devices to the cloud. Our future work will identify the specific causes of the long shutdown times and alleviate any problems in the infrastructure to reduce the shutdown time of the simulation. For instance, we will try whether different packet routing policies or network configurations might help mitigate such problems. In conclusion, our study demonstrates the impact of communication dataflows on the battery consumption of IoT devices in vehicles, highlighting a significant correlation that is crucial for understanding overall energy efficiency in VANETs. The findings validate the effectiveness of SimulatorBridger in providing accurate simulations of VANET environments based on IoT infrastructure. Future research could explore traffic rerouting strategies based on battery optimization criteria, leveraging the capabilities of our proposed simulator.

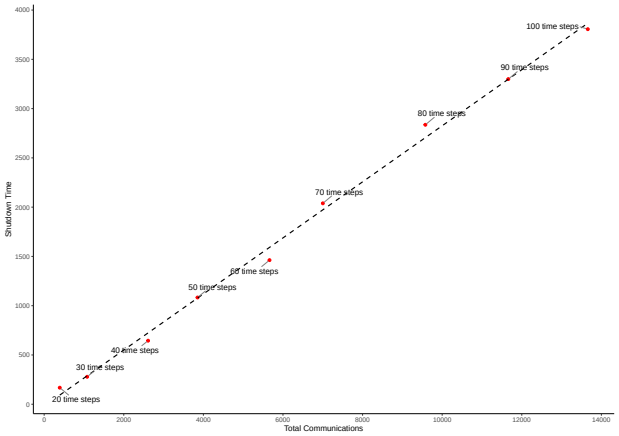


Figure 9: Correlation between simulation time, number of undergoing communications, and their effect to the shutdown time.

Conflict of Interest The authors declare no conflict of interest.

References

[1] B. Boukenadil, "Importance of realistic mobility models for vanet network simulation", *arXiv preprint arXiv:1410.2450*, 2014.

[2] J. Harri, F. Filali, C. Bonnet, "Mobility models for vehicular ad hoc networks: a survey and taxonomy", *IEEE Communications Surveys & Tutorials*, vol. 11, no. 4, pp. 19–41, 2009, doi:10.1109/SURV.2009.090403.

[3] T. Szydlo, A. Szabala, N. Kordiumov, K. Siuzdak, L. Wolski, K. Alwasel, F. Habeeb, R. Ranjan, "Iotsim-osmosis-res: Towards autonomic renewable energy-aware osmotic computing", *Software: Practice and Experience*, vol. 52, no. 7, pp. 1698–1716, 2022, doi:10.1002/spe.3084.

[4] P. Á. López, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, E. WieBner, "Microscopic traffic simulation using SUMO", "ITSC", pp. 2575–2582, IEEE, 2018, doi:10.1109/ITSC.2018.8569938.

[5] A. Dahlinger, B. Ryder, F. Wortmann, "Car as a sensor. paying people for providing their car data", "Proceedings of the 5th International Conference on Internet of Things, Seoul, South Korea", 2015, doi:10.3929/ethz-b-000106154.

[6] R. A. Eaton, R. H. Joubert, E. A. Wright, "Pothole primer – a public administrator's guide to understanding and managing the pothole problem", Special Report Vol. 81–21, US Army Corps of Engineers – Cold Regions Research & Engineering Laboratory, 1989.

[7] M. H. Asad, S. Khaliq, M. H. Yousaf, M. O. Ullah, A. Ahmad, "Pothole detection using deep learning: A real-time and ai-on-the-edge perspective", *Advances in Civil Engineering*, vol. 2022, p. 9221211, 2022, doi:10.1155/2022/9221211.

[8] M. Saleem, S. Abbas, T. M. Ghazal, M. Adnan Khan, N. Sahawneh, M. Ahmad, "Smart cities: Fusion-based intelligent traffic congestion control system for vehicular networks using machine learning techniques", *Egyptian Informatics Journal*, 2022, doi:10.1016/j.eij.2022.03.003.

[9] N. Baranasuriya, V. Navda, V. N. Padmanabhan, S. Gilbert, "Qprobe: Locating the bottleneck in cellular communication", "CoNEXT", 2015, doi:10.1145/2716281.2836118.

[10] J. Kang, X. Chen, D. Wu, Y. T. Xu, X. Liu, G. Dudek, T. Lee, I. Park, "Hierarchical policy learning for hybrid communication load balancing", "ICC 2021 - IEEE Int. Conf. on Communications", pp. 1–6, 2021, doi:10.1109/ICC42927.2021.9500379.

[11] M. Villari, M. Fazio, S. Dustdar, O. Rana, R. Ranjan, "Osmotic computing: A new paradigm for edge/cloud integration", *IEEE Cloud Computing*, vol. 3, no. 06, pp. 76–83, 2016, doi:10.1109/MCC.2016.124.

[12] K. Alwasel, D. N. Jha, E. Hernandez, D. Puthal, M. Barika, B. Varghese, S. K. Garg, P. James, A. Zomaya, G. Morgan, *et al.*, "Iotsim-sdwan: A simulation framework for interconnecting distributed datacenters over software-defined wide area network (sd-wan)", *Journal of Parallel and Distributed Computing*, vol. 143, pp. 17–35, 2020, doi:10.1016/j.jpdc.2020.04.006.

[13] S. Häger, S. Böcker, C. Wietfeld, "3d self-motion tracking services: Coalescence of mmWave beam orientations and phase information", "2022 IEEE FNWF", 2022, doi:10.1109/FNWF55208.2022.00009.

[14] L. Bieker, D. Krajzewicz, A. Morra, C. Michelacci, F. Cartolano, "Traffic simulation for all: a real world traffic scenario from the city of bologna", "Modeling Mobility with Open Data: 2nd SUMO Conference 2014 Berlin, Germany, May 15-16, 2014", pp. 47–60, Springer, 2015, doi:10.1007/978-3-319-15024-6_4.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Received: 10 May 2024, Revised: 17 June 2024, Accepted: 18 June 2024, Online: 13 July 2024

DOI: https://doi.org/10.55708/js0307002

A Thorough Examination of the Importance of Machine Learning and Deep Learning Methodologies in the Realm of Cybersecurity: An Exhaustive Analysis

Ramsha Khalid *[1,2] **, Muhammad Naqi Raza** [1]

[1] Electrical Engineering Technology, University of Sialkot, Sialkot, 51310, Pakistan

[2] Electrical Engineering, University of Lahore, Lahore, 53720, Pakistan

*Corresponding author: Ramsha Khalid, University of Sialkot, Sialkot, Email: ramshakhalid2404@gmail.com

ABSTRACT: In today's digital age, individuals extensively engage with virtual environments hosting a plethora of public and private services alongside social platforms. As a consequence, safeguarding these environments from potential cyber threats such as data breaches and system disruptions becomes paramount. Cybersecurity encompasses a suite of technical, organizational, and managerial measures aimed at thwarting unauthorized access or misuse of electronic information and communication systems. Its objectives include ensuring operational continuity, safeguarding the confidentiality and integrity of sensitive data, and shielding consumers from various forms of cyber intrusions. This paper delves into the realm of cybersecurity practices devised to fortify computer systems against diverse threats including hacking and data breaches. It examines the pivotal role of artificial intelligence within this domain, offering insights into the utilization of machine learning and deep learning techniques. Moreover, it synthesizes key findings from relevant literature exploring the efficacy and impact of these advanced methodologies in cybersecurity. Findings underscore the substantial contributions of machine learning and deep learning techniques in fortifying computer systems against unauthorized access and mitigating the risks posed by malicious software. These methodologies facilitate proactive measures by predicting and comprehending the behavioral patterns and traffic associated with potential cyber threats.

KEYWORDS: Cybersecurity, Cyber attackers, Artificial intelligence, Machine learning, Deep learning, Communication systems, Unauthorized entry

1. Introduction

The Internet has transformed various facets of contemporary life, creating a global village where knowledge exchange and cultural interactions flourish. Networks form the backbone of this digital landscape, connecting devices such as computers and mobile phones. Their significance lies in enabling access to the Internet; without it, these devices lose much of their functionality. Through networks, data, information, and applications flow seamlessly via physical cables and wireless radio waves. Protecting personal data, crucial for transactions conducted over networks, is paramount amidst the looming threat of hackers aiming for identity theft or fraudulent activities [1]. The COVID-19 pandemic catalyzed a shift towards digital transactions with minimal physical contact to mitigate virus transmission. This transition propelled widespread adoption of electronic transactions by institutions and businesses, highlighting their efficiency and accessibility benefits for consumers. Simultaneously, online shopping platforms, including those on social media like Facebook, witnessed increased activity, facilitating the sale and distribution of goods [2,3]. Moreover, educational institutions transitioned to online platforms for delivering education and training, reflecting a growing acceptance of digital learning modalities. Similarly, remote work gained traction as a viable option for both public and private sector organizations, facilitated by the widespread availability of Internet connectivity [4,5].

As remote work liberates employees from fixed locations, the sharing of online workspaces necessitates information security specialists to evaluate the associated business risks and prevent unauthorized access or hacking attempts from external parties [6,7]. Despite the implementation of sophisticated technical security measures by organizations to combat cyber threats, the human element remains a critical consideration, as employees' skills often constitute the weakest link in the security chain. It is imperative for employees to be vigilant against potential hacking or malicious software threats that may compromise their data unbeknownst to them. In addition to conducting awareness-raising activities such as training sessions and workshops for employees lacking expertise in cybersecurity, organizations must also enforce a range of technical safeguards. Practices posing threats to information security include leaving computers unlocked while unattended, abandoning devices in public settings, and disregarding company policies regarding password security. Consequently, there is a pressing need for further investigation into the risks associated with remote work.

1.1. Artificial Intelligence

Artificial intelligence techniques have emerged as some of the most advanced and invaluable tools in various fields, including cybersecurity and information security [8,9]. AI encompasses the capability of machines, electronic devices, software, applications, and gaming consoles to mimic human brain functions, such as awareness, memory, and data utilization, in decision-making processes [10]. Equipped with electronic brains, AI-enabled devices can analyze data and perform required operations, leveraging insights garnered from experimental data. The term "cybersecurity" has gained prominence in response to the widespread adoption and accessibility of Internet networks, particularly with the advent of 5G technology [11]. The proliferation of electronic crimes targeting data, information, and applications on computers and electronic devices underscores the imperative for robust security measures. Consequently, companies are increasingly turning to AI-based techniques to forecast cybercrime activities, preempt attacks, and thwart unauthorized intrusions into computer systems. Compared to human specialists, AI techniques offer superior efficacy in scrutinizing network users for authorization, thereby enhancing security protocols [12,13]. Moreover, their capacity for rapid learning, retention, and task execution translates into significant time and resource savings for experts. Notably, AI techniques excel in recognizing repetitive patterns, a feature invaluable in cybersecurity for identifying and analyzing user behaviors and predicting anomalous activities indicative of malware infiltration [14,15].

This paper makes a substantial contribution b elucidating the pivotal role of machine learning and deep learning techniques in cybersecurity. It showcases their efficacy in mitigating intrusions and attacks on computer systems while elucidating their diverse applications within the cyber domain. Additionally, the paper provides a succinct overview of seminal studies leveraging these techniques in cybersecurity, scrutinizing their findings and elucidating their impact on decision-making processes. The data utilized in this paper are sourced from reputable news outlets and scholarly literature, streamlining the research process for cybersecurity scholars and practitioners.

The subsequent sections of this paper are structured as follows: Section 2 provides an examination of prevailing cybersecurity practices and the attendant challenges confronting computer systems. In Section 3, an overview of prominent datasets employed in attack and intrusion detection is presented. Sections 4 and 5 delve into the importance of machine learning and deep learning methodologies in cybersecurity, along with a comprehensive review of key literature utilizing these techniques. Finally, Section 6 offers concluding remarks.

2. Cybersecurity Practices

In recent years, the electronics and technology industry has experienced significant growth, becoming an integral part of daily life for individuals, indispensable for the fulfillment of business endeavors and projects. The functionality of modern devices relies on a suite of applications designed to serve human needs, necessitating comprehensive protective measures to safeguard against intrusions, hacking, attacks, and unauthorized access [16,17]. Concerns regarding hacking and data theft loom large for numerous companies and institutions. As organizations across diverse sectors increasingly acknowledge the paramount importance of their data, attention to cybersecurity has surged. This encompasses various facets, including measures to secure communication systems, data, and raw information, as well as virtual and physical components associated with operating systems. Secure applications, accessible only to authorized personnel, are vital elements within this framework [18–20]. Described as a combination of tools and practices, cybersecurity serves to defend computer systems' contents and thwart the infiltration of malicious software [21]. Table 1 provides a comprehensive summary of the various types of cybersecurity and their roles in safeguarding computer systems. Fundamental to cybersecurity are three key features: confidentiality, ensuring that unauthorized individuals cannot access or manipulate data within a computer system; integrity, preventing unauthorized modification or deletion of data; and availability, ensuring that data, information, and communications reach intended recipients without interception or decryption by unauthorized parties. Regardless of location, cyberattacks pose significant risks

to organizations, their employees, and their clientele, potentially resulting in profound consequences. Thus, it is imperative for employees to possess awareness of their organizations' cybersecurity protocols and adopt practices to mitigate associated risks. Table 2 illustrates significant instances of cyberattacks.

Table 1: Types of Cybersecurity and Their Functions

Cybersecurity Category	Functions
Application Security	Execute intricate codes to safeguard and encrypt data effectively [22]
Information Security	Ensure data protection from unauthorized access and alterations [23]
Infrastructure Security	Secure critical infrastructures such as power networks and data centers, ensuring absence of vulnerabilities [24]
Network Security	Protect networks from intrusions using tools like remote access management, two-factor authentication (2FA), and robust firewalls [25]
User Education	Provide valuable training sessions and conferences for employees and cybersecurity professionals [26]

Table 2: Types of Cybersecurity Attacks

Type	Description
Malware	A collection of malicious applications designed to damage systems and steal data [27]
Ransomware	Malicious software that encrypts data, disables systems, and restricts authorized user access [28]
Phishing	A common form of social engineering wherein individuals are manipulated into divulging sensitive information, posing significant security risks [29]
DDoS	Denial-of-Service attacks that disrupt systems, preventing user access to network resources, and inflicting financial or reputational harm on organizations [30]
SQL Injection	Exploits web security vulnerabilities to access, steal, modify, or delete data from websites, leading to system dysfunction [31]
Zero-Day Exploit	Newly discovered security vulnerabilities exploited by hackers to target computer systems, often leaving administrators with insufficient time to address the issue [32]
DNS Tunnelling	A sophisticated attack technique involving the encoding of system data and applications, challenging to detect [33]
XSS Attacks	Injection of malware into trusted websites, camouflaged as benign browser scripts [34]

The initial instance of malicious software, Creeper, surfaced in the 1970s with the capacity to destruct computer data. Upon infection, affected computers displayed a notable message on their screens: "I'm a creeper, catch me if you can!". In response to this emerging threat, the inaugural antivirus program, known as Reaper, was developed.

In 1903, Nevil Maskelyne made history as the first documented hacker by intercepting the inaugural wireless telegraph transmission, thereby revealing vulnerabilities inherent in Marconi's system. Concurrently, John Draper emerged as the pioneering cybercriminal, having discovered that the whistle included in Cap'n Crunch cereal boxes emitted a tone capable of deceiving telephone exchange signals, enabling him to place unauthorized free calls..

2.1. Cybersecurity Data Science

Data science encompasses the analysis of diverse data domains, ranging from life sciences to consumer behavior and cybersecurity. It plays a pivotal role in shaping the future of systems and cybersecurity fields, given its reliance on comprehensive data sets. Detecting cyber threats hinges on the meticulous analysis of security data, including files, records, and user activities within a network. Cybersecurity professionals leverage various techniques such as file hashes and custom rules, such as signatures or heuristics, to trace the origins of incoming data streams. While these manual methods offer unique advantages, they demand significant efforts to stay abreast of evolving threats and breaches.

Figure 1 illustrates the transformation of big data into actionable decisions, while Table 3 delineates various types of cybersecurity attacks. Data science endeavors to revolutionize information technology by leveraging machine learning and deep learning techniques to identify and address system vulnerabilities through feature extraction and pattern recognition from training data. Over the past decade, cybersecurity has increasingly relied on data science and artificial intelligence due to their capability to convert raw data into actionable insights and fortify system security.

In essence, data science offers an efficient approach to decision-making through tasks such as data engineering for data accumulation and analysis, data volume reduction through critical data filtering, discovery of unique patterns and data learning techniques, development of innovative data-driven security models, knowledge generation for mitigating false alerts, and optimization of system resources.

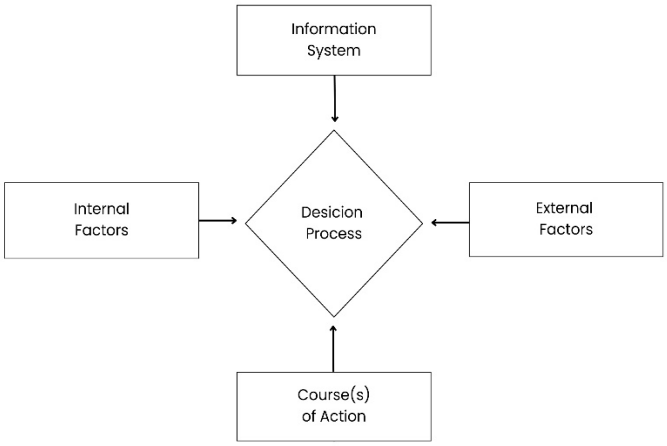


Figure 1: Representation of the process of data analysis and decision-making

Table 3: Prominent datasets utilized in cybersecurity research.

Dataset	Description
DARPA	Contains intrusion detection data, including LLDOS-1.0 and LLDOS-2.0.2, depicting connections between source and destination IP addresses, categorized by MIT Lincoln Laboratory for evaluating attacks and intrusion detection [35].
CAIDA	Encompasses distributed denial of service (DDoS) attack traffic and regular traffic traces, including unspecified traffic from a 2007 DDoS attack, facilitating evaluation of machine-learning-based detection models for identifying DoS activity on the Internet [36].
CTU-13	Comprises botnet traffic captured by a Czech university in 2011, featuring real botnet traffic mixed with normal and background traffic across 13 scenarios, suitable for data-based malware analysis employing machine learning techniques [37].
KDD'99 Cup	Widely used dataset since 1999 with 41 features for evaluating anomaly detection, categorizing attacks into probing, remote-to-local (R2L), user-to-remote (U2R), and DoS, suitable for assessing machine-learning-based attack detection models [38].
NSL-KDD	Revised version of KDD'99 Cup dataset, eliminating redundant records and addressing inherent issues to avoid bias towards frequent records [39].
MAWI	Dataset aiding researchers in anomaly detection, sourced from Japanese network research institutions, featuring traffic deviation labels in the MAWI archive, revised daily to include all traffic from applications and malware [40].
ISCX'12	Produced by Canadian Institute for Cybersecurity, containing 19 features for machine-learning-based attack detection and network penetration models, used in real-time with expert input to prevent system destruction and data theft [41].
Bot-IoT	Simulated Internet of Things environment dataset from UNSW Canberra Cyber Range Lab, featuring reliable traffic and various attack types, including DDoS, DoS, OS and service scan, keylogging, and data exfiltration, organized by protocol [42].
ISOT'10	Mix of malicious and non-malicious data traffic dataset from University of Victoria's ISOT research, utilized for evaluating models, machine-learning-based classification, and attack and penetration localization [43].
UNSW-NB15	Created using IXIA PerfectStorm tool in UNSW Canberra Cyber Range Lab, comprising contemporary synthetic attack activities and behaviors, with 49 features and 9 attack types, generated from TCPDUMP, ARGUS, and Bro-IDS tools [44].

2.2. Machine Learning in Cybersecurity

Electronic devices are experiencing rapid advancements, attracting a substantial following across various domains. However, the extensive communication and data exchange among these devices pose significant risks, notably concerning data breaches. Scholars advocate for leveraging machine learning techniques to

mitigate electronic threats, although these methods are still in developmental stages. Cyber threats are dynamic, necessitating adaptive solutions. Machine learning stands out as a potent tool due to its adaptability and learning capabilities. While effective in detecting and thwarting known malware attacks, machine learning faces challenges against novel threats. Operating within the realm of artificial intelligence, machine learning employs statistical operations to analyze data, extract insights, and aid decision- making. Its core objective is to enable computers to learn from expert-provided data [45–47]. Machine learning techniques encompass various rules and methodologies aimed at identifying or predicting novel data patterns or behaviors. These techniques find application in cybersecurity and can be categorized into supervised and unsupervised methods. Despite the substantial adoption of machine learning in cybersecurity, these tools remain imperfect, demanding significant human oversight, and necessitating continuous retraining of algorithms due to the inability to fully automate data processes [48,49]. This segment explores the functionality of machine learning techniques and their integration into cybersecurity practices. Figure 2 elucidates the operational framework of these techniques in detecting anomalies within a system.

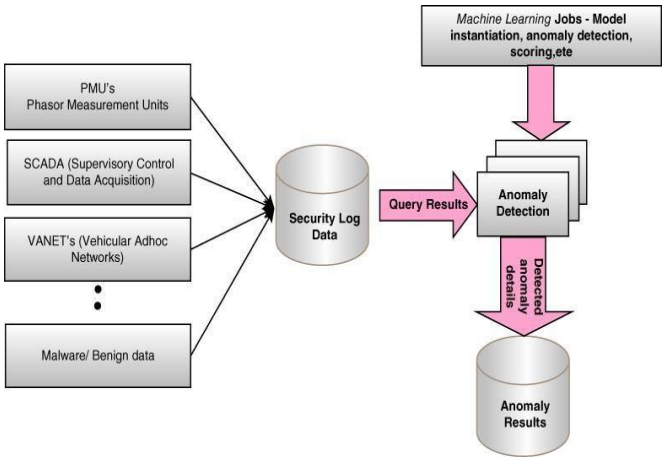


Figure 2: Utilizing machine learning techniques for anomaly detection [54].

2.2.1. *Supervised Learning*

Supervised learning functions methodically by establishing precise objectives and achieving them through a defined set of inputs [50]. Widely employed across various domains, supervised learning techniques offer straightforward implementation and monitoring. They are typically classified into classification and regression methods, which respectively categorize security data or anticipate specific security issues to be addressed in the future. The failure of an organization to avert an anticipated attack on its computer system can have far-reaching consequences, leading to substantial financial losses and necessitating a laborious recovery process in Figure 3. Consequently, the utilization of machine learning techniques in cybersecurity is advocated to bolster data protection across all sectors and safeguard the data of both organizations and their users. Notable supervised learning techniques in classification include logistic regression, decision trees, support vector machines, k-nearest neighbors, and naive Bayes. These techniques are also applied in prediction tasks owing to their capacity to construct data-driven predictive models. For instance, the activities of users within a network, whether in a public or private institution, can be forecasted by continually tracking their actions, gathering relevant data, and discerning between processes initiated by human users and those generated by bots impacting the network. Meanwhile, prominent regression techniques in supervised learning, such as linear regression and support vector regression, are utilized to identify underlying causes of significant cybercrimes that profoundly affect individuals' lives and develop corresponding solutions. The distinction between classification and regression techniques lies in their respective outcomes: classification yields categorical or discrete results/effects, whereas regression produces numeric or continuous outputs/effects.

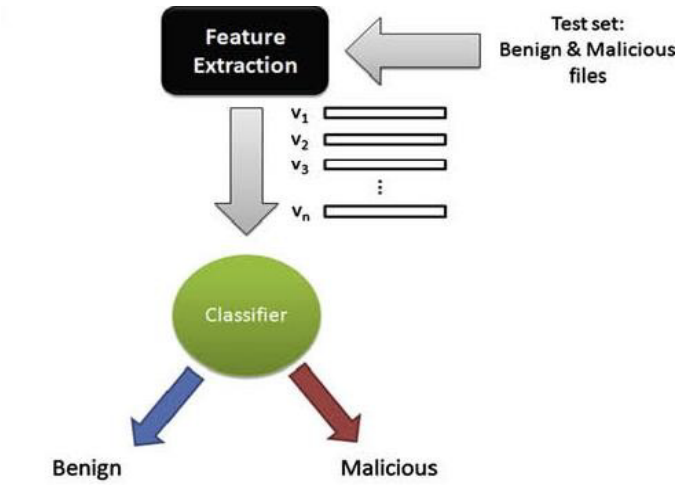


Figure 3: Utilizing machine learning techniques to classify unidentified datasets as either malicious or benign.

2.2.2. *Unsupervised Learning*

The primary objective of unsupervised learning techniques is to unveil patterns, structures, or insights from unlabelled data. However, within cybersecurity, malware often evade detection by dynamically altering their operations [51]. Clustering techniques, including k-means, k- medoids, and single linkage, constitute unsupervised learning methods aimed at uncovering hidden and intricate attack patterns and structures within large datasets. These techniques play a pivotal role in identifying and notifying users or developers about anomalies within systems, breaches of privacy policies, and unauthorized data accesses. Engineering tasks associated with these technologies, such as optimizing dataset features or extracting pertinent features related to specific security issues, are deemed essential for

conducting further analyses, irrespective of dataset scale. Moreover, security features are prioritized based on their significance. Additional methods like linear discriminant analysis, principal component analysis, non-negative matrix factorization, and Pearson correlation analysis contribute to addressing cybersecurity threats and uncovering clandestine programs using machine learning to preempt attacks and data breaches. In expert systems, rules are manually defined and implemented by a knowledge engineer in collaboration with a cybersecurity expert. Association rules learning aims to identify existing rules or relationships among datasets to extract relevant security attributes. Correlation analysis assesses the strength of relationships among datasets. Data mining techniques are categorized into frequent pattern-based, logic-based, and tree-based methods. Techniques such as AIS, Apriori, Apriori-TID, Apriori-Hybrid, FP-Tree, RARM, and Eclat are employed to formulate association rules capable of detecting intrusion and data theft issues. Table 4 enumerates the ten most impactful studies concerning the application of machine learning techniques in identifying attacks on operating systems.

Table 4: Scholarly works exploring the application of machine learning methodologies in identifying attacks and malicious software.

Article	Purpose	Techniques	Most Suitable Effect
[52]	Detect distributed denial- of-service (DDoS) attacksand malicious data.	MLP, K-NN, SVM, FL, ED, and MNB	The most suitable classification method is the MLP technique, achieving an F1-score of over 98% for emulated traffic and over 99% for real traffic.
[53]	Develop an intrusion detection system employing an efficient and reliable classifier.	SVM	The authors achieve a high accuracy of 98.62%,deemed excellent in intrusion detection.
[54]	Design a knowledge-based alert verification strategy with an intelligent filter to eliminate unwanted alarms.	KNN	KNN demonstrates the highest performance accuracy of 93.2% and the most robust F-measure of 91.8%.
[55]	Conduct experiments on five million Android applications to detect malicious software in the Android OS by recognizing features.	DL, FFC, Y-MLP, and DT	These methods achieve apeak performance accuracy of over 98% in identifying malicious software.
[56]	Detect ransomware tools (RANDS) operating within Windows environments through three stages (ransomware analysis, learning, and detection).	NB and DT	These methods attain an average classification accuracy of 96.27% in categorizing ransomware, with a 1.32% average real-time execution error.
[57]	Utilize a dataset of 300,000 attributes to predict and detect ransomware.	SVM	The technique reports an accuracy exceeding 88% in ransomware classification.
[58]	Monitor systems and detect intrusions by analyzing incoming data activities of servers and identifying malicious software.	DT with binary split	This technique achieves an impressive attack detection accuracy of over 99%.
[59]	Enhance the performance of the random forest strategy to detect misuse, anomaly, and hybrid-	New systematic frameworks of RF	These frameworks achieve a high detection rate in identifying and reporting anomalies.

	network-based intrusion detection systems (IDSs).		
[60]	Detect denial-of-service (DOS) attacks in software-defined networks (SDNs) and address cybersecurity management in SDN architectures.	KNN, RF, SVM, and ANN	These methods achieve an accuracy exceeding 98% in detecting DOS attacks.
[61]	Detect intrusions and identify malicious data through data mining techniques.	FCM, ANN, and SVM	These methods achieve a peak accuracy of 98.99% in detecting remote to-local (R2L) attacks.

However, machine learning techniques are not without their constraints. For instance, they are incapable of identifying attacks that have not been previously encountered. Furthermore, the detection of behavioral patterns and anomalies may result in false positives if the behavioral restriction policy is overly broad. Conversely, implementing a stricter policy could diminish the effectiveness of these techniques. The selection of appropriate datasets is also crucial during the training phase of machine learning techniques in cybersecurity endeavors. Without proper training, these techniques may fail to deliver the anticipated results. When cyber adversaries become aware of a security system relying solely on a single defense technology, they may devise strategies to bypass such systems, such as through hacking. Nonetheless, a robust cybersecurity framework grounded in machine learning can leverage multiple complementary techniques.

2.3. Deep Learning in Cybersecurity

To address the intricate challenges in cybersecurity, various methodologies are employed based on specific criteria such as data volume, issue type, sensitivity, and decision tolerance. Deep learning techniques, leveraging parallel processing, prove highly effective in handling large- scale data and entail complex procedures [62–64]. This section critically examines the literature implementing deep learning methodologies for intrusion detection, attack mitigation, and malware identification. These studies are succinctly outlined in Table 5. Deep learning architectures are not configured solely on local platforms; rather, they are deployed on server- based systems to ensure data integrity, confidentiality, and reliability, while also safeguarding against unauthorized access. The development of a robust deep learning model in cybersecurity involves two primary stages. Initially, the data transfer area is encrypted within the local environment before transmitting to the server. Subsequently, upon reaching the server, these encrypted data are processed, classified, and identified. For example, in character recognition from images, the first stage encompasses character encoding and transmission, while the second stage involves processing the received data and detecting any potential man-in-the- middle attacks between the server and local system, crucial for accurate data classification. This approach ensures secure information transmission to users, preventing unauthorized access to the system.

In networked environments, ensuring appropriate security measures and robust data preservation, storage, and transmission mechanisms are imperative responsibilities of the companies managing these systems. Breaches in networked systems vary depending on network activity and scale, with larger and more active networks encountering higher volumes of data requiring processing. To efficiently handle such data influx, parallel processing and deep learning techniques are favored for their speed and accuracy [65–67]. Deep learning methodologies have emerged as effective tools for detecting malware, given their capability to analyze various characteristics of malicious programs that can disrupt system operations by altering data. Numerous researchers have employed convolutional neural networks for classifying data, extracting essential features, and isolating genetic sequences indicative of malicious applications, thereby facilitating network training. Moreover, deep learning techniques are instrumental in identifying biological attributes such as personal identification numbers (PINs) and passwords, recognizing user voices or images, and analyzing behavior- based licenses. At this stage, techniques derived from recurrent neural networks (RNNs), including gated recurrent units and long short-term memory, are often deployed.

Device security stands as a pivotal concern within the realm of cybersecurity, wherein heightened security measures necessitate increased interactions between humans and electronic environments. Deep learning techniques play a crucial role in safeguarding data, systems, and applications. Renowned for their exceptional performance in processing 2D and 3D media data as well as vast datasets, these techniques are

extensively employed in image and video processing. In the cybersecurity domain, deep learning endeavors to discern the suitability of received data for supervised or unsupervised techniques and to evaluate the influence of prior knowledge on subsequent insights. Moreover, deep learning assesses system performance in addressing problems across both one- and multi-dimensional examples. Scholars are actively exploring deep learning methodologies to devise solutions for numerous cybersecurity challenges [68–70].

Table 5: Scholarly works employing deep learning methodologies for the detection of attacks and malicious software.

Article	Purpose	Technique use	Most Suitable Effect
[71]	Protect autonomous vehicle systems from attacks and ensure control.	CNN and CNN-LSTM	Achieves a remarkably high accuracy exceeding 97% in identifying attack messages and preventing their display on vehicle screens.
[72]	Real-time intrusion detection in vehicular data encompassing cyber and physical processes.	LR, SVM, RF, DT, MLP, and RNN	RNN exhibits the best accuracy of 79.3% in detecting malware, denial-of-service (DoS) attacks, and command injections.
[73]	Predict software vulnerabilities and identify accessible features at an early stage.	ExBERT framework	ExBERT framework reveals over 46,000 vulnerabilities with a prediction accuracy surpassing 91%.
[74]	Analyze and detect attacks using URLs on edge devices, safeguarding data in cloud-Internet of Things (IoT) systems.	Multiple concurrent deep models	These models achieve a high accuracy of over 99% in detecting normal requests.
[75]	Develop an intrusion detection application and protect computer systems.	MLP and PID	MLP and PID attain an accuracy of 98.96% in intrusion detection and understanding attack types.
[76]	Detect intrusions and analyze network anomalies.	K-NN and DNN	DNN achieves over 92% accuracy in intrusion detection.
[77]	Establish an intrusion detection approach for cyber-attack security and classification.	RNN	RNN achieves the highest accuracy of 98.27%.
[78]	Enhance simulation training methods to detect anomaly intrusion via the Internet.	RBM and DBM	RBM and DBM demonstrate an exceptional accuracy of 97.9% in detecting anomaly intrusions.
[79]	Identify malicious programs in multi-cloud healthcare systems using the MUSE model.	DHSNN	DHSNN achieves excellent training and testing accuracies ranging from 95% to 100% in detecting new attacks on dataflows.
[80]	Classify a traffic detection system and network fault identification or intrusion detection system.	CNN	CNN reports an efficacy of over 99%.

3. Conclusion and Future Work

Artificial intelligence stands as a cornerstone of the Fourth Industrial Revolution, poised to continue its profound impact on society due to its manifold benefits. However, to strike a harmonious balance between technological advancement and fundamental human values, the nexus between artificial intelligence and cybersecurity requires thorough exploration. It is imperative to harness modern technologies within virtual environments and social networking platforms to safeguard user privacy and information. The evolution of artificial intelligence capabilities must parallel the emergence of novel applications, fostering digital collaboration among nations and leveraging the integration of digital technologies into physical settings. Facilitating unrestricted access to data for researchers, while upholding user privacy, is essential for training artificial intelligence algorithms and conducting data analysis on a broader scale. Increased financial and ethical investments in machine learning and deep learning are vital to fortify the privacy of social media users and mitigate data breaches. Continuous training for cybersecurity professionals on cutting-edge technologies, hacker tactics, and malware behavior is imperative. Furthermore, stringent penalties and fines should be enforced for the misuse of artificial intelligence techniques and unauthorized privacy breaches. Future research endeavors should focus on implementing these techniques for predictive and classification purposes in cybersecurity, thus optimizing their efficacy.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] N. Bhalaji, "Reliable Data Transmission with Heightened Confidentiality and Integrity in IOT Empowered Mobile Networks," *Journal of IoT in Social, Mobile, Analytics, and Cloud*, vol. 2, no. 2, 106–117, 2020, doi:10.36548/jismac.2020.2.004.

[2] J. Budd, B.S. Miller, E.M. Manning, V. Lampos, M.Z. et al., "Digital technologies in the public-health response to COVID-19," *Nature Medicine*, vol. 26, 1183–1192, 2020, doi:10.1038/s41591-020-1011-4.

[3] K. Leung, J.T. Wu, G.M. Leung, "Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing," *Nature Communications*, vol. 12, no. 1501, 1–8, 2021, doi:10.1038/s41467-021-21776-2.

[4] S. Shrestha, S. Haque, S. Dawadi, R.A. Giri, "Preparations for and practices of online education during the Covid-19 pandemic: A study of Bangladesh and Nepal," *Education and Information Technologies*, vol. 27, 243–265, 2021, doi:10.1007/s10639-021-10659-0.

[5] M. Ssenyonga, "Imperatives for post COVID-19 recovery of Indonesia's education, labor, and SME sectors," *Cogent Economics & Finance*, vol. 9, no. 1, 1–51, 2021, doi:10.1080/23322039.2021.1911439.

[6] H. Saleous, M. Ismail, S.H. AlDaajeh, N. Madathil, S. Alrabaee, "COVID-19 pandemic and the cyberthreat landscape: Research challenges and opportunities," *Digital Communications and Networks*, vol. In press, , 2022, doi:10.1016/j.dcan.2022.06.005.

[7] H.S. Lallie, L.A. Shepherd, J.R.C. Nurse, A. Erola, G.E. et al., "Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic," *Computers & Security*, vol. 105, 102248, 2021, doi:10.1016/j.cose.2021.102248.

[8] J. Li, "Cyber security meets artificial intelligence: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, 1462–1474, 2019, doi:10.1631/FITEE.1800573.

[9] Z. Zhang, H. Ning, F. Shi, F. Farha, Y. Xu, F.Z. et al., "Artificial intelligence in cyber security: research advances, challenges, and opportunities," *Artificial Intelligence Review*, vol. 55, 1029–1053, 2021, doi:10.1007/s10462-021-09976-0.

[10] M.M. Mijwil, "Implementation of Machine Learning Techniques for the Classification of Lung X-Ray Images Used to Detect COVID-19 in Humans," *Iraqi Journal of Science*, vol. 62, no. 6, 2099–2109, 2021, doi:10.24996/ijs.2021.62.6.35.

[11] J. Cáceres-Hidalgo, D. Avila-Pesantez, "Cybersecurity Study in 5G Network Slicing Technology: A Systematic Mapping Review," in *Proceedings of IEEE Fifth Ecuador Technical Chapters Meeting*, IEEE, Cuenca, Ecuador: 1–6, 2021, doi:10.1109/ETCM53643.2021.9590742.

[12] T. Ghosh, H. Al Banna, S. Rahman, S. Kaiser, M.M. et al., "Artificial intelligence and internet of things in screening and management of autism spectrum disorder," *Sustainable Cities and Society*, vol. 74, 103189, 2021, doi:10.1016/j.scs.2021.103189.

[13] A. Adadi, M. Lahmer, S. Nasiri, "Artificial Intelligence and COVID-19: A Systematic umbrella review and roads ahead," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, 5898–5920, 2022, doi:10.1016/j.jksuci.2021.07.010.

[14] M. Abdullahi, Y. Baashar, H. Alhussian, A. Alwadain, N.A. et al., "Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review," *Electronics*, vol. 11, no. 2, 1–27, 2022, doi:10.3390/electronics11020198.

[15] I.F. Kilincer, F. Ertam, A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," *Computer Networks*, vol. 188, 107840, 2021, doi:10.1016/j.comnet.2021.107840.

[16] S. Kuipers, M. Schonheit, "Data Breaches and Effective Crisis Communication: A Comparative Analysis of Corporate Reputational Crises," *Corporate Reputation Review*, vol. 25, 176–197, 2021, doi:10.1057/s41299-021-00121-9.

[17] N. Rawindaran, A. Jayal, E. Prakash, C. Hewage, "Cost Benefits of Using Machine Learning Features in NIDS for Cyber Security in UK Small Medium Enterprises (SME)," *Future Internet*, vol. 13, no. 8, 1–36, 2021, doi:10.3390/fi13080186.

[18] F. Quayyum, D.S. Cruzes, L. Jaccheri, "Cybersecurity awareness for children: A systematic literature review," *International Journal of Child-Computer Interaction*, vol. 30, 100343, 2021, doi:10.1016/j.ijcci.2021.100343.

[19] P. Formosa, M. Wilson, D. Richards, "A principlist framework for cybersecurity ethics," *Computers & Security*, vol. 109, 102382, 2021, doi:10.1016/j.cose.2021.102382.

[20] I.H. Sarker, H. Furhad, R. Nowrozy, "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions," *SN Computer Science*, vol. 2, no. 173, 2021, doi:10.1007/s42979-021-00557-0.

[21] E. Fosch-Villaronga, T. Mahler, "Cybersecurity, safety and robots: Strengthening the link between cybersecurity and safety in the context of care robots," *Computer Law & Security Review*, vol. 41, 105528, 2021, doi:10.1016/j.clsr.2021.105528.

[22] P. Sharma, S. Jain, S. Gupta, V. Chamola, "Role of machine learning and deep learning in securing 5G-driven industrial IoT applications," *Ad Hoc Networks*, vol. 123, 102685, 2021, doi:10.1016/j.adhoc.2021.102685.

[23] A. Rehman, T. Saba, T. Mahmood, Z. Mehmood, M.S. et al., "Data hiding technique in steganography for information security using number theory," *Journal of Information Science*, vol. 45, no. 6, 767–778, 2018, doi:10.1177/0165551518816303.

[24] G. Hale, C. Bartlett, "Managing the Regulatory Tangle: Critical Infrastructure Security and Distributed Governance in Alberta's Major Traded Sectors," *Journal of Borderlands Studies*, vol. 34, no. 2, 257–279, 2018, doi:10.1080/08865655.2017.1367710.

[25] Y. Wang, A. Smahi, H. Zhang, H. Li, "Towards Double Defense Network Security Based on Multi-Identifier Network Architecture," *Sensors*, vol. 22, no. 3, 1–17, 2022, doi:10.3390/s22030747.

[26] D.G. Broo, U. Boman, M. Törngren, "Cyber-physical systems research and education in 2030: Scenarios and strategies," *Journal of Industrial Information Integration*, vol. 21, 100192, 2021, doi:10.1016/j.jii.2020.100192.

[27] M.M. Mijwil, "Malware Detection in Android OS Using Machine Learning Techniques," *Data Science and Applications*, vol. 3, no. 2, 5–9, 2020.

[28] U. Urooj, B.A.S. Al-rimy, A. Zainal, F.A. Ghaleb, M.A. Rassam, "Ransomware Detection Using the Dynamic Analysis and Machine Learning: A Survey and Research Directions," *Applied Sciences*, vol. 12, no. 1, 1–45, 2021, doi:10.3390/app12010172.

[29] A.F. AL-Otaibi, E.S. Alsuwat, "A Study on Social Engineering Attacks: Phishing Attack," *International Journal of Recent Advances in Multidisciplinary Research*, vol. 7, no. 11, 6374–6379, 2020.

[30] A. Narote, V. Zutshi, A. Potdar, R. Vichare, "Detection of DDoS Attacks using Concepts of Machine Learning," *International Journal for Research in Applied Science & Engineering Technology*, vol. 10, no. VI, 390–403, 2022.

[31] N. Bedeković, L. Havaš, T. Horvat, D. Crčić, "The Importance of Developing Preventive Techniques for SQL Injection Attacks," *Tehnički Glasnik*, vol. 16, no. 4, 523–529, 2022, doi:10.31803/tg-20211203090618.

[32] U.K. Singh, C. Joshi, D. Kanellopoulos, "A framework for zero-day vulnerabilities detection and prioritization," *Journal of Information Security and Applications*, vol. 46, 164–172, 2019, doi:10.1016/j.jisa.2019.03.011.

[33] Y. Wang, A. Zhou, S. Liao, R. Zheng, R. Hu, L. Zhang, "A comprehensive survey on DNS tunnel detection," *Computer Networks*, vol. 179, 108322, 2021, doi:10.1016/j.comnet.2021.108322.

[34] Y. Zhou, P. Wang, "An ensemble learning approach for XSS attack detection with domain knowledge and threat intelligence," *Computers & Security*, vol. 82, 261–269, 2019, doi:10.1016/j.cose.2018.12.016.

[35] J. He, C. Chang, P. He, M.S. Pathan, "Network Forensics Method Based on Evidence Graph and Vulnerability Reasoning," *Future Internet*, vol. 8, no. 4, 1–18, 2016, doi:10.3390/fi8040054.

[36] M.P. Singh, A. Bhandari, "New-flow based DDoS attacks in SDN: Taxonomy, rationales, and research challenges," *Computer Communications*, vol. 15, 509–527, 2020, doi:10.1016/j.comcom.2020.02.085.

[37] J.L.G. Torres, C.A. Catania, E. Veas, "Active learning approach to label network traffic datasets," *Journal of Information Security and Applications*, vol. 49, 102388, 2019, doi:10.1016/j.jisa.2019.102388.

[38] S. Choudhary, N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT," *Procedia Computer Science*, vol. 167, 1561–1573, 2020, doi:10.1016/j.procs.2020.03.367.

[39] L. Dhanabal, S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, 446–452, 2015.

[40] B. Bouyeddou, F. Harrou, B. Kadri, Y. Sun, "Detecting network cyber-attacks using an integrated statistical approach," *Cluster Computing*, vol. 24, 1435–1453, 2020, doi:10.1007/s10586-020-03203-1.

[41] M. Idhammad, K. Afdel, M. Belouch, "Semi-supervised machine learning approach for DDoS detection," *Applied Intelligence*, vol. 48, 3193–3208, 2018, doi:10.1007/s10489-018-1141-2.

[42] N. Koroniotis, N. Moustafa, E. Sitnikova, B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, 779–796, 2019, doi:10.1016/j.future.2019.05.041.

[43] I.H. Sarker, "Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective," *SN Computer Science*, vol. 2, no. 154, 1–16, 2021, doi:10.1007/s42979-021-00535-6.

[44] S.M. Kasongo, Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset," *Journal of Big Data*, vol. 7, no. 105, 1–20, 2020, doi:10.1186/s40537-020-00379-6.

[45] R.T. S., R. Sathya, "Ensemble Machine Learning Techniques for Attack Prediction in NIDS Environment," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 2, 78–82, 2022, doi:10.52866/ijcsm.2022.02.01.008.

[46] Y. Niu, A. Korneev, "Identification Method of Power Internet Attack Information Based on Machine Learning," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 2, 1–7, 2022, doi:10.52866/ijcsm.2022.02.01.001.

[47] M.M. Mijwil, E.A. Al-Zubaidi, "Medical Image Classification for Coronavirus Disease (COVID-19) Using Convolutional Neural Networks," *Iraqi Journal of Science*, vol. 62, no. 8, 2740–2747, 2021, doi:10.24996/ijs.2021.62.8.27.

[48] M. Sarhan, S. Layeghy, N. Moustafa, M. Gallagher, M. Portmann, "Feature extraction for machine learning-based intrusion detection in IoT networks," *Digital Communications and Networks*, vol. In press, , 2022, doi:10.1016/j.dcan.2022.08.012.

[49] M.A. Teixeira, T. Salman, M. Zolanvari, R. Jain, N. Meskin, M. Samaka, "SCADA System Testbed for Cybersecurity Research Using Machine Learning Approach," *Future Internet*, vol. 10, no. 8, 1–15, 2018, doi:10.3390/fi10080076.

[50] K. Aggarwal, M.M. Mijwil, Sonia, A.H. Al-Mistarehi, S. Alomari, M. Gök, A.M. Alaabdin, S.H. Abdulrhman, "Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning," *Iraqi Journal for Computer Science and Mathematics*, vol. 3, no. 1, 115–123, 2022, doi:10.52866/ijcsm.2022.01.01.013.

[51] L.F. Maimó, A.H. Celdrán, A.L.P. Gómez, F.J.G. Clemente, J. Weimer, I. Lee, "Intelligent and Dynamic Ransomware Spread Detection and Mitigation in Integrated Clinical Environments," *Sensors*, vol. 19, no. 5, 1–31, 2019, doi:10.3390/s19051114.

[52] V.M. Rios, P.R.M. Inácio, D. Magoni, M.M. Freire, "Detection of reduction-of-quality DDoS attacks using Fuzzy Logic and machine learning algorithms," *Computer Networks*, vol. 186, 107792, 2021, doi:10.1016/j.comnet.2020.107792.

[53] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," *Expert Systems with Applications*, vol. 39, no. 1, 424–430, 2012, doi:10.1016/j.eswa.2011.07.032.

[54] W. Meng, W. Li, L. Kwok, "Design of intelligent KNN-based

alarm filter using knowledge-based alert verification in intrusion detection," *Security and Communication Networks*, vol. 8, no. 18, 3883–3895, 2015, doi:10.1002/sec.1307.

[55] A. Mahindru, A.L. Sangal, "MLDroid—framework for Android malware detection using machine learning techniques," *Neural Computing and Applications*, vol. 33, 5183–5240, 2020, doi:10.1007/s00521-020-05309-4.

[56] H. Zuhair, A. Selamat, "RANDS: A Machine Learning-Based Anti-Ransomware Tool for Windows Platforms," in *Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques*, 573–587, 2019, doi:10.3233/FAIA190081.

[57] U. Adamu, I. Awan, "Ransomware Prediction Using Supervised Learning Algorithms," in *Proceedings of International Conference on Future Internet of Things and Cloud*, Istanbul, Turkey: 1–6, 2019, doi:10.1109/FiCloud.2019.00016.

[58] S. Puthran, K. Shah, "Intrusion Detection Using Improved Decision Tree Algorithm with Binary and Quad Split," in *Proceedings of International Symposium on Security in Computing and Communication*, 427–438, 2016, doi:10.1007/978-981-10-2738-3_37.

[59] J. Zhang, M. Zulkernine, A. Haque, "Random-Forests-Based Network Intrusion Detection Systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 38, no. 5, 649–659, 2008, doi:10.1109/TSMCC.2008.923876.

[60] F. Musumeci, A.C. Fidanci, F. Paolucci, F. Cugini, M. Tornatore, "Machine-Learning-Enabled DDoS Attacks Detection in P4 Programmable Networks," *Journal of Network and Systems Management*, vol. 30, no. 21, 2021, doi:10.1007/s10922-021-09633-5.

[61] A.M. Chandrasekhar, K. Raghuveer, "Confederation of FCM clustering, ANN and SVM techniques to implement hybrid NIDS using corrected KDD cup 99 dataset," in *Proceedings of International Conference on Communication and Signal Processing*, Melmaruvathur, India: 1–6, 2014, doi:10.1109/ICCSP.2014.6949927.

[62] S. Ahmed, Z.A. Abbood, H.M. Farhan, B.T. Yasen, M.R. Ahmed, A.D. Duru, "Speaker Identification Model Based on Deep Neural Networks," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 1, 108–114, 2022, doi:10.52866/ijcsm.2022.01.01.012.

[63] A.K. Faieq, M.M. Mijwil, "Prediction of Heart Diseases Utilising Support Vector Machine and Artificial Neural Network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, 374–380, 2022, doi:10.11591/ijeecs.v26.i1.pp374-380.

[64] M.M. Mijwil, R.A. Abttan, A. Alkhazraji, "Artificial intelligence for COVID-19: A Short Article," *Asian Journal of Pharmacy, Nursing and Medical Sciences*, vol. 10, no. 1, 1–6, 2022, doi:10.24203/ajpnms.v10i1.6961.

[65] K. Shaukat, S. Luo, V. Varadharajan, I.A. Hameed, S. Chen, et al., "Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity," *Energies*, vol. 13, no. 10, 1–27, 2020, doi:10.3390/en13102509.

[66] D. Chen, P. Wawrzynski, Z. Lv, "Cyber security in smart cities: A review of deep learning-based applications and case studies," *Sustainable Cities and Society*, vol. 66, 102655, 2021, doi:10.1016/j.scs.2020.102655.

[67] P. Suresh, K. Logeswaran, R.M. Devi, K. Sentamilselvan, G.K. Kamalam, H. Muthukrishnan, Contemporary survey on effectiveness of machine and deep learning techniques for cyber security, 177–200, 2022, doi:10.1016/B978-0-323-85209-8.00007-9.

[68] M. Taseer, H. Ghafory, "SQL Injection Attack Detection Using Machine Learning Algorithm," *Mesopotamian Journal of CyberSecurity*, 5–17, 2022, doi:10.58496/MJCS/2022/002.

[69] I.E. Salem, M. Mijwil, A.W. Abdulqader, M.M. Ismaeel, A. Alkhazraji, A.M.Z. Alaabdin, "Introduction to The Data Mining Techniques in Cybersecurity," *Mesopotamian Journal of CyberSecurity*, 28–37, 2022, doi:10.58496/MJCS/2022/004.

[70] R.T. Rasheed, Y. Niu, S.N. Abd, "Harmony Search for Security Enhancement," *Mesopotamian Journal of CyberSecurity*, 5–8, 2021, doi:10.58496/MJCS/2021/002.

[71] T.H.H. Aldhyani, H. Alkahtani, "Attacks to Automatous Vehicles: A Deep Learning Algorithm for Cybersecurity," *Sensors*, vol. 22, no. 1, 1–20, 2022, doi:10.3390/s22010360.

[72] G. Loukas, T. Vuong, R. Heartfield, G. Sakellari, Y. Yoon, et al., "Cloud-Based Cyber-Physical Intrusion Detection for Vehicles Using Deep Learning," *IEEE Access*, vol. 6, 3491–3508, 2017, doi:10.1109/ACCESS.2017.2782159.

[73] J. Yin, M. Tang, J. Cao, H. Wang, "Apply transfer learning to cybersecurity: Predicting exploitability of vulnerabilities by description," *Knowledge-Based Systems*, vol. 210, 106529, 2020, doi:10.1016/j.knosys.2020.106529.

[74] Z. Tian, C. Luo, J. Qiu, X. Du, M. Guizani, "A Distributed Deep Learning System for Web Attack Detection on Edge Devices," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, 1963–1971, 2020, doi:10.1109/TII.2019.2938778.

[75] A. Thirumalairaj, M. Jeyakarthic, "Perimeter Intrusion Detection with Multi Layer Perception using Quantum Classifier," in *Proceedings of International Conference on Inventive Systems and Control*, Coimbatore, India: 1–6, 2020, doi:10.1109/ICISC47916.2020.9171159.

[76] K. Atefi, H. Hashim, M. Kassim, "Anomaly Analysis for the Classification Purpose of Intrusion Detection System with K-Nearest Neighbors and Deep Neural Network," in *Proceedings of Conference on Systems, Process and Control*, Melaka, Malaysia: 1–6, 2019, doi:10.1109/ICSPC47137.2019.9068081.

[77] M. Almiani, A. AbuGhazleh, A. Al-Rahayfeh, S. Atiewi, A. Razaque, "Deep recurrent neural network for IoT intrusion detection system," *Simulation Modelling Practice and Theory*, vol. 101, 102031, 2020, doi:10.1016/j.simpat.2019.102031.

[78] K. Alrawashdeh, C. Purdy, "Toward an Online Anomaly Intrusion Detection System Based on Deep Learning," in *Proceedings of International Conference on Machine Learning and Applications*, Anaheim, CA, USA: 1–6, 2016, doi:10.1109/ICMLA.2016.0040.

[79] L. Gupta, T. Salman, A. Ghubaish, D. Unal, A.K. Al-Ali, R. Jain, "Cybersecurity of multi-cloud healthcare systems: A hierarchical deep learning approach," *Applied Soft Computing*, vol. 118, 108439, 2022, doi:10.1016/j.asoc.2022.108439.

[80] W. Wang, M. Zhu, X. Zeng, X. Ye, Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *Proceedings of International Conference on Information Networking*, Da Nang, Vietnam: 1–6, 2017, doi:10.1109/ICOIN.2017.7899588.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Engr. Ramsha Khalid has done her bachelor's degree from Lahore College for Women University, Lahore in 2018. She has done her master's degree from University of Lahore, Lahore in 2022. She is working as Lecturer at University of Sialkot since 2019. Her area of interest includes Computer & Communication Networks, Machine Learning, Artificial Intelligence,

Cyber Security, Control Systems and Renewable Energy Systems. Recently she has published a conference paper in IEEE INMIC'23 held in University of Central Punjab, Lahore as a first author.



Engr. M. Naqi Raza has done his bachelor's degree from University of Gujrat, Gujrat in 2018. He has done his master's degree from University of Sialkot, Sialkot in 2024.

He is working as Junior Lecturer at University of Sialkot since 2019. His area of interest includes Power Generation (Conventional and Renewable), Wind Power Generation and Utilization, Optimization of Wind Energy, Solar Energy, Solar Power Applications, Electric Vehicles (PHEVs).

Received: 16 May 2024, Revised: 02 July 2024, Accepted: 09 July 2024, Online: 23 July 2024

DOI: https://doi.org/10.55708/js0307005

A Computational Approach for Recognizing Text in Digital and Natural Frames

Mithun Dutta[1], Dhonita Tripura[1], Jugal Krishna Das[2]

[1]Department of Computer Science and Engineering, Rangamati Science and Technology University, Rangamati-4500, Bangladesh

[2]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

*Corresponding author: Mithun Dutta, mithundutta92@gmail.com

ABSTRACT: Acquiring tenable text detection and recognition outcomes for natural scene images as well as for digital frames is very challenging emulating task. This research approaches a method of text identification for the English language which has advanced significantly, there are particular difficulties when applying these methods to languages such as Bengali because of variations in script, morphology. Text identification and recognition is accomplished on multifarious distinct steps. Firstly, a photo is taken with the help of a device and then, Connected Component Analysis (CCA) and Conditional Random Field (CRF) model are introduced for localization of text components. Secondly, a merged model (region-based Convolutional Neural Network (Mask-R-CNN) and Feature Pyramid Network (FPN)) are used to detect and classify text from images into computerized form. Further, we introduce a combined method of Convolutional Recurrent Neural Network (CRNN), Connectionist Temporal Classification (CTC) with K-Nearest Neighbors (KNN) Algorithm for extracting text from images/ frames. As the goal of this research is to detect and recognize the text using a machine learning-based model a new Fast Iterative Nearest Neighbor (Fast INN) algorithm is now proposed based on patterns and shapes of text components. Our research focuses on a bilingual issue (Bengali and English) as well as it producing satisfactory image experimental outcome with better accuracy and it gives around 98% accuracy for our proposed text recognition methods which is better than the previous studies.

KEYWORDS: Iterative, Component, Recognition, Bilingual

1. Introduction

Recently, a notable surge in research efforts dedicated to text recognition, a critical aspect within various image processing and vision algorithms. Scene text recognition, in particular, emerges as a challenging yet highly advantageous endeavor involving the identification of text within natural images. Deciphering text embedded in digital photographs poses a significant challenge. Moreover, image text recognition not only forms the foundation of information retrieval but also plays a crucial role in enabling effective human-machine communication. The potential applications of a system capable of identifying and extracting text from real-world images are extensive. Additionally, databases comprise texts of diverse types, including manually edited caption texts and scene texts with various orientations, thereby introducing layers of complexity to text detection and recognition. Notably, viewers often prioritize text when interpreting images, emphasizing the significance of text detection and recognition in aiding human comprehension of intricate visual compositions.

Numerous methodologies have been developed to address the challenges associated with text detection and recognition in digital frames and natural scene images characterized by varying orientations, scripts, font sizes, and other factors. Previous studies suggest that many existing methods focus on specific data types and address particular issues such as complex backgrounds, low contrast, or multiple scripts and orientations, resulting in suboptimal performance when confronted with data influenced by multiple adverse factors [1, 2, 3]. The primary challenges stem from various sources: frames captured with low-resolution cameras which make low contrast inputs; images also taken with high-resolution devices may exhibit high contrast input but frequently feature complex backgrounds, leading to an increased incidence of false positives; images sourced

from digital platforms often showcase a multitude of character components.

In this study, we employed mask-RCNN for character and text detection and recognition of both Bangla and English sentences. R-CNN (Region-based Convolutional Neural Network), including its variant Mask R-CNN, represents a class of machine learning models tailored explicitly for computer vision tasks, particularly object detection. Mask R-CNN propagates Faster R-CNN by introducing an additional pulse for outputting object masks alongside class labels and bounding-box offsets, thereby facilitating the extraction of finer spatial layouts of objects. Additionally, our proposed method incorporates the Fast Iterative Nearest Neighbor (Fast INN) algorithm, which utilizes shape information to detect candidate components. Subsequently, Fast INN extracts shape of script. The novelty of our approach is to explore these fundamental concepts to address the ongoing challenges of text detection and recognition without imposing rigid constraints, thereby aiming for enhanced performance and versatility.

2. Literature Review

In [1], the authors proposed a method for detecting and tracking text on a variety of large and small text blocks. This is consists of two different modules: first one is a sum of squared difference (SSD) and another one is a contour-based module. The main purpose of the paper proposed [2] text recognition, which submitted a petition on the text lines detection.

In [3, 4], the authors proposed the text detection method which is carried out through edge detection, local thresholding, and hysteresis edge recovery. In [5, 6], the researches proposed an SVM classifier which used to identify text from the selected features object. In [7, 8], the authors addressed a model that is based on invariant features. In [9, 10], the authors proposed signal extraction techniques based on RWT method for extracting test information. They also proposed text line extraction based on integrated K-shortest path optimization. In [11, 12], the authors used the NLP to analyze and become acquainted the data set and also used both Naive Bayes and logistic regression algorithms to determine the best accuracy system.

3. Proposed System

3.1. Proposed Framework

In this section, we outline the operational procedures of the proposed machine learning-based text recognition framework and delve into its logical execution. The experiment integrates a preprocessing technique tailored to mitigate unwanted noise and artifacts effectively. Initially, we employ the Connected Component Analysis (CCA) model to determine adjacent pixels based on predefined pixel connectivity criteria. Additionally, we utilize edge detection methodologies. Typically, false values in the image are associated with background pixels, while valid values represent foreground or object pixels. The connected component analysis methodology segments the image into regions, identifies relevant areas, and extracts text from these regions. Subsequently, the image is partitioned into smaller components, which are then classified based on their geometric attributes.

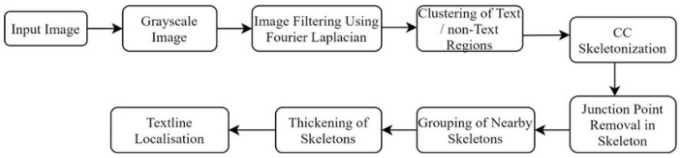


Figure 1: Text localization based on connected component analysis

Text localization involves the utilization of conditional random field (CRF), a statistical modeling technique employed when class labels for different inputs which are independent. Specifically, in image segmentation, the assignment of a class label to a pixel is contingent upon the labels of its neighboring pixels. To calculate the confidence of region texts, translate the output, conditioned variable (s), s ∈ {acc, rej}, and t can be calculated based on the Bayes' theorem, where Pt (s|x) are calculated is as follows.

$$P_t(x|s) = \frac{P_t(s|x)P_t(x)}{\sum_x^1 P_t(s|x)P_t(x)} = \frac{P_t(s|x)P_{t-1}(x|accept)}{\sum_x^1 P_t(s|x)P_{t-1}(x|accept)}$$

Image Segmentation: Niblack's binarization formula [13] is defined as:

$$b(x) = \begin{cases} 0 & , \ if\ gray(x) < \mu_r(x) - k \cdot \sigma_r(x); \\ 255 & , \ if\ gray(x) > \mu_r(x) + k \cdot \sigma_r(x); \\ 100 & , \ other, \end{cases}$$

And an implementation, P (Y |X) is as:

$$P(Y|X) = \frac{1}{Z(X)} \exp\big(-E(X, Y, N, \lambda)\big),$$

Where Z(X) is the normalization constant which is used to solve the problem, and the best label Y ∗ might be considered by maximizing conditional probability P (Y |X) to minimize the energy.

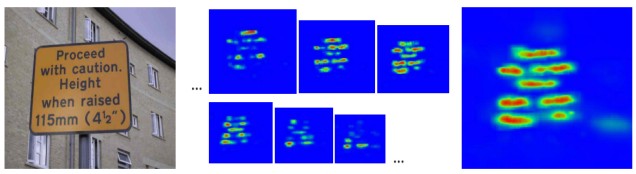


Figure 2: (a) original image (b) text confidence maps (c) text confidence map for the original image

Then, Mask R-CNN is an advance computer vision model used for object instance segmentation. A text detector based on Mask R-CNN is used, and fully convolutional

networks mainly inspire the methods. First, CNN is adopted to detect text blocks from which character candidates are extracted.

Then, FPN (Feature Pyramid Network) is used to predict the corresponding segmentation masks. Convolutional Neural Network (CNN) is a one kind of AI based neural network which uses for recognizing and processing image components that is optimized to process pixel data. It integrates data components detection task where the goal is to detect object through the formation of bounding box prediction of an image and a semantic segmentation task, that may classify every pixel into pre-defined object categories.

This is an execution of Mask R-CNN on Python 3, and the model generates bounding boxes and segmentation masks for each data object in the frame, which is also based on FPN.

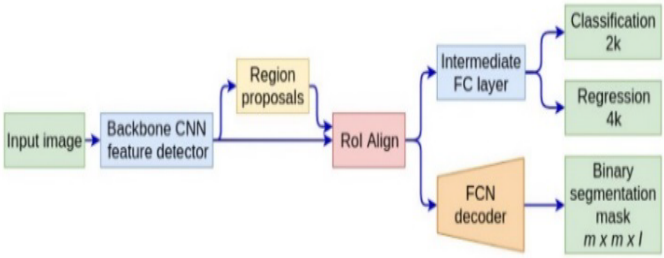


Figure 3: Flowchart of segmentation with Mask R-CNN

Mask R-CNN combines object detection and instance segmentation by using a Feature Pyramid Network (FPN) and the Region of Interest Align (ROIAlign) layer, Mask R-CNN which achieves strong performance and accuracy rate.

Now, a combined method of Convolutional Recurrent Neural Network (CRNN), Connectionist Temporal Classification (CTC) with K-Nearest Neighbors (KNN) Algorithm for extracting text from images. Extracting text of different shapes and sizes, various directions and orientations from images, especially from web pages and sites which is augmented reality assistance systems, and content moderation in social media platforms with the combined method. Combining Convolutional Recurrent Neural Networks (CRNN) with Connectionist Temporal Classification (CTC) allows CRNN to handle variable-length sequences without requiring an explicit alignment between input images and text outputs.

Suppose a given sequence $X = [x1, x2, x3, x4, x5\ldots, xT]$, such as any voice, and by mapping the sequence to its corresponding outcome sequence $Y = [y1, y2, y3, y4, y5\ldots,yU]$. Our main object is to find out an accurate mapping between $X's$ and $Y's$.

In particular:

- X and Y are different in length or dimension.
- The lengths ratio of X and Y may vary from each to other.
- No accurate alignment of X and Y.

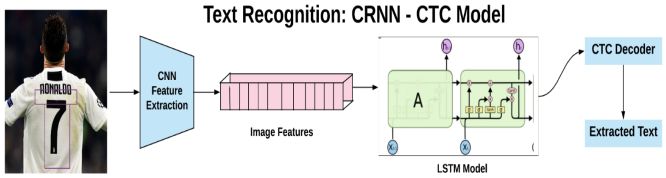


Figure 4: Text Recognition Pipeline Architecture

In CTC, a conditional probability $p(\pi|x)$ can be labeled as follows:

$$p(\pi|x) = \prod_{t=1}^{T} y_{\pi_t}^t, \forall \pi \epsilon L'^T$$

$$p(s|x) = \sum_{\pi \epsilon B^{-1}(8)}^{1} p(\pi|x)$$

CTC loss function is described as

$$L_{CTC} = -\ln p(s|x)$$

Further, the k-nearest neighbor (KNN) classifier uses proximity to make classifications or predictions about the grouping of an individual data point and fundamentally relies on a distance metric. The most common choice is the Minkowski distance dist(x,z)=(d Σ r=1|xr−zr|p)1/p. Finally, the proposed method, Fast Iterative Nearest Neighbor algorithm (Fast INN), is applied to recognize the text. The FINN algorithm is suggested by the inspiration of KNN, which is one of the simplest Machine Learning algorithms. The proposed algorithm estimates the similarity between the new case/data and the available cases and places the new case in the category that is most similar to the available data. During the training phase, the FINN algorithm stores the datasets, and when it gets new data, it then classifies that data into a category that is similar to the latest data. As the FINN algorithm helps to identify the nearest points or the groups for a query point and to determine the closest groups or the nearest points for a query point, we need to calculate some distance metrics (Euclidean distance, Manhattan distance, Minkowski distance, etc.).

Our Fast Iterative Nearest Neighbor (Fast I-NN) working can be explained through the following algorithm:

Step-1: Select the number of the neighbors (Suppose the number is N)

Step-2: Calculate the distance (Euclidean) of N number of neighbors

Step-3: Take the N nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these N neighbors (adjacent data points), count the number of the data points in each category.

Step-5: Assign the neoteric data points to that category

Step-6: Find for which the number of the neighbor is maximum.

Step-7: Calculate the distance is minimum.

Step-8: Compare the new data pixel with residual data points to that category.

Step-9: Predict and Print the output.

3.2. *Datasets*

In this study, research will use most of the original dataset images, which are taken with a Kodak DX7590 (5.0 MP) still frame camera and a Sony DCR-SR85E handy camera in still mode, as well as from Kaggle. From different background images are taken like roadways, digital banner, signboard, wall writings, etc. Data may be prepared for a machine learning approach using Python language through Jupyter Notebook. The proposed text recognition method is admired in both cases qualitatively and quantitively. Datasets include original images of dimension either (321x481) or (481x321) pixels, and from the datasets randomly, 80 percent of images are used for training and 20 percent of images for testing. For Augmentation, we used to tilt the image with different angle, zooming, rotation, shearing width and height shifting, and horizontal and vertical flipping techniques in a Python setup environment.

3.3. *Configure and Train the Model*

The model, Default-Trainer, Default-Predictor, Color-Mode, and Visualizer were all imported. We also imported mask-RCNN with FPN, as well as other models and the associated checkpoints. We defined the dataset and other settings, such as total number of workers, batch size, as well as the number of classes, by importing Default-Trainer. We trained the model further after initializing it using pre-trained weights. We've trained our new model in Jupyter Notebook and Google Colab wherein it could dash and make advantages. The system generated an output folder, into which the trainer saved the training checkpoints to count the maximum and best checkpoints for the trained model. The trained software automatically saves the checkpoints in the output directory during training.

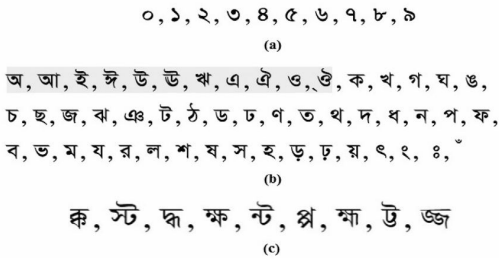


Figure 5: Bangla character (a) Digits (b) Vowels and consonants (c) Compound/joint characters



Figure 6: Frequently used grapheme roots of Bangla Char

3.4. *Eliminate unwanted Character*

This system eliminates multiple-identified characters from the object array after replacement. Machine learning detection models can predict the same object or predict the same instance with a different class at another time. Consequently, we must remove redundant objects from the array, keeping only one instance wherein the same instance is not recognized more than once.

3.5. *Recognize the Text*

All the images previous step is used for both to train and test [9, 10]. Our proposed method has been accustomed for text line recognition. The following figure shows the step of capturing to recognize.



Figure 7: From image capture to text recognition

4. Experimental Result

4.1. *Performance Evaluation*

Maximum accuracy rate indicates the perfect prediction for the localization and recognition of text. A

minimal false positive rate indicates fewer redundant pixels in unwanted regions resulting in a higher accuracy rate. This provides a balanced measure of the model's ability to correctly identify both positive and negative instances. In this paper, we have adapted the performance of machine learning models. The figure below reincarnates the results of the trained machine learning. Here, Figure a. shows accuracy rate of performance, and Fig b. shows false negative prediction, Fig c. indicates accuracy of the foreground classification, Fig d. shows the accuracy rate of mask RCNN, Fig e. represents the false negative of mask RCNN, and lastly, Fig f. shows the total loss of the mask RCNN.

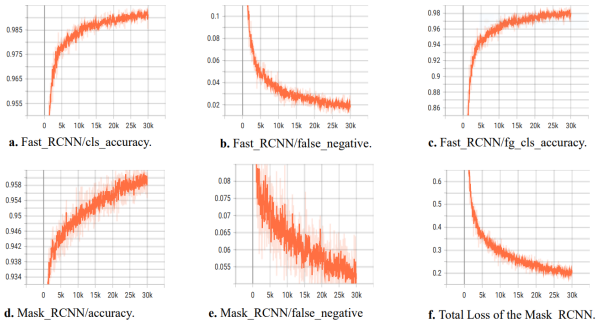


Figure 8: Result of the Trained Machine Learning Model

4.2. Experimental Result and Discussion

Our proposed model for future bilingualism text recognition will perform very well. We have used a huge number of data set to acquire a better performance and a well predicted result. To achieve the performance and print the result we had to perform our operational work properly in several sectors from where quite a few data were lost. Using the proposed algorithm through python language significant percentage accuracy rate is achieved. To calculate the sample size and accuracy rate we need the total number of successfully recognized images. The following table depicts the results of our study of all steps together.

Table 1: Summarizes the accuracy in terms of its gradations

Gradations	Text Detection	Text Extraction	Text Recognition
Total Data Samples	19200	19200	19200
Correct for Bangla	18430	18110	17998
Accuracy for Bangla	95.98%	94.32%	93.74%
Correct for English	19110	19030	18810
Accuracy for English	99.53%	99.11%	97.97%

5. Conclusion

In this research, we have well-acquainted and analyzed various text detection techniques for natural scenes as well as for digital images. In here, we introduced a unique model for detecting and recognizing the Bangla and English text using machine learning and the proposed Fast Iterative Nearest Neighbor (FINN) method. Before the system detects text, images are initially taken from the nature or surrounding environment and from that image texts are extracted. After that, text recognition was done by a merge proposed method where it gave false positive results or was unable to recognize some words or parts of them [13]. Then, the proposed method works effectively on curved or slanted Bengali and English texts as well. Pre-processing techniques make this model more robust against any random artifacts or unwanted blurring effects. However, the system gives about 98% accuracy for the proposed text recognition model.

References

[1] H. Li, D. Doermann, and O. Kia, "Automatic Text Detection and Tracking in Digital Video," *IEEE Transactions on Image Processing*, vol. 9, no. 1, 147-156, Jan. 2000, doi: 10.1109/83.817607.

[2] D. Chen, J.-M. Odobez, and H. Bourlard, "Text detection and recognition in images and video frames," *Pattern Recognition*, vol. 37, 595-608, 2004, doi:10.1016/j.patcog.2003.06.001.

[3] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A Laplacian Approach to Multi-Oriented Text Detection in Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, 412-420, Feb. 2011, doi:10.1109/TPAMI.2010.166.

[4] M. R. Lyu, J. Song, and M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, 243-255, Feb. 2005, doi:10.1109/TCSVT.2004.841653.

[5] X.-C. Yin, Z.-Y. Zuo, S. Tian, and C.-L. Liu, "Text Detection, Tracking and Recognition in Video: A Comprehensive Survey," *IEEE Transactions on Image Processing*, 1-24, 2015, doi: 10.1109/TIP.2016.2554321.

[6] Q. Ye, Q. Huang, W. Gao, and D. Zhao, "Fast and robust text detection in images and video frames," *Image and Vision Computing*, vol. 23, 565-576, 2005, doi:10.1016/j.imavis.2005.01.004.

[7] K. S. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal, and T. Lu, "Multi-Script-Oriented Text Detection and Recognition in Video/Scene/Born Digital Images," *IEEE Transactions on Circuits and Systems for Video Technology*, doi: 10.1109/TCSVT.2018.2817642, 2018.

[8] M. Cai, J. Song, and M. R. Lyu, "A New Approach for Video Text Detection," in *Proceedings. International Conference on Image Processing*, 2002, 117-120, doi: 10.1109/ICIP.2002.1037973.

[9] C. S. Shin, K. I. Kim, M. H. Park, and H. J. Kim, "Support Vector Machine-Based Text Detection in Digital Video," in *Proceedings of IEEE ICIP*, 2000, 634-641.

[10] H. Wang, S. Huang, and L. Jin, "Focus On Scene Text Using Deep Reinforcement Learning," in *Proceedings of the 24th International Conference on Pattern Recognition (ICPR)*, Beijing, China, Aug. 2018, 3759-3766, doi: 10.1109/ICPR.2018.8545022.

[11] Y. Wang, "Extraction Algorithm of English Text Information from

Color Images Based on Radial Wavelet Transform," *Special Section on Gigapixel Panoramic Video with Virtual Reality*, Aug. 2020, doi 10.1109/ACCESS.2020.3020621.

[12] O. Y. Ling, L. B. Theng, A. Chai, and C. McCarthy, "A Model for Automatic Recognition of Vertical Text in Natural Scene Images," in *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, Penang, Malaysia, 2018, doi: 10.1109/ICCSCE.2018.8685019.

[13] M. Dutta, A. Mohajon, S. Dev, D. S. Bappi, and J. K. Das, "Text Recognition of Bangla and English Scripts in Natural Scene Images," *International Journal of Advanced Research in Science and Technology*, vol. 12, no. 10, 1137-1142, 2023.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Received: 12 July, 2024, Revised: 20 July, 2024, Accepted: 21 July, 2024, Online: 01 August, 2024
DOI: https://dx.doi.org/10.55708/js0308001

Dynamic and Partial Grading of SQL Queries

Benard Wanjiru*, Patrick van Bommel, Djoerd Hiemstra
Radboud University, Nijmegen, Postbus 9010 6500 GL Nijmegen, The Netherlands
*Corresponding author: Benard Wanjiru, email: benard.wanjiru@ru.nl

ABSTRACT: Automated grading systems can help save a lot of time when evaluating students' assignments. In this paper we present our ongoing work for a model for generating correctness levels. We utilize this model to demonstrate how we can grade students SQL queries employing partial grading in order to allocate points to parts of the queries well written and to enable provision of feedback for the missing parts. Furthermore, we show how we can grade the queries taking into account the skill level of students at different stages of SQL learning process. We divide the stages into introductory, intermediary, and advanced stages and in each apply different type of grading that takes account the students' knowledge at that stage. We implemented this model in our class and graded 5 quizzes containing more than 25 different questions for 309 students. We discuss 3 examples for each stage and offer comprehensive examples of the model in action.

KEYWORDS: Correctness Levels, Software Correctness, Automated Grading, Assessment, Partial Marks, SQL Query Grading

1. Introduction

Partial grading in SQL software exercises is awarding a selected portion or a fraction of a whole grade. It involves giving points to parts whereby the student has done well or deducting points to parts missed. An advantage of partial grading is that we are able to acknowledge students' efforts by awarding points to the correct parts of a query. This helps in protecting their motivation to learn [1]. Furthermore, we are able to pinpoint the parts of a query the student has missed. This helps in providing constructive feedback to help the student improve on their work. Partial grading systems which are able to carry out this form of grading have been extensively researched [2, 3, 4, 5]. Nevertheless, these systems do not address different skill levels of students at various stages of the learning process. We propose a model in which students' skill levels are taken into consideration. We refer to this as *dynamic grading*.

There are three types of knowledge acquired by students in programming courses, syntactic, conceptual, and strategic [6]. In our project, we broadly classify these groups of knowledge into 2 types, syntactic and semantic [7, 8] as it makes it manageable to translate our grading model into a running program. Students learn well when taught syntax and semantic knowledge in parallel while syntax knowledge develops as a result on repetition [8]. Even though this is the case, novice students might struggle more with syntax during the introductory lessons than at the end [9]. Various methods have been proposed of how to teach SQL in various steps. Some research proposes teaching SQL in an increasing order of difficulty [10]. This order begins from teaching 'SELECT' syntax to recursive SQL at the end. Other research proposes teaching SQL in a series of steps involving building a query one clause at a time [11]. The first clause would be SELECT followed by FROM clause then WHERE clause and so on. A mental model in which learners learn SQL starting with remembering SQL concepts followed by comprehending the concepts with creating the queries at the end have been proposed as well [12]. In the various learning styles proposed, students are novice during the start of the class struggling with syntax and at the end have accumulated enough knowledge to write completely semantically correct queries. We considered to utilize these findings to grading as well. To this effect, our goals in this paper are to demonstrate how we can:

- Assign a discrete level that shows how correct a query is, in relation to an instructor's specifications.
- Factor in students' knowledge level when calculating grades.

Also, to demonstrate whether:

- It is possible for an automated grading tool to match the grades given by manual graders.
- If students consider an automated grading tool to be fair.

In this context, a fair grade is a grade that matches the one that would be given by an instructor. Furthermore, this grading capability should be applied consistently to other answers as well.

To factor in students' knowledge level, we divided a full programming course's time into three stages. introductory stage, intermediate stage, and the advanced stage. We used syntax, semantics, and results as features for evaluation. For each stage, these three features which we refer to as

properties were evaluated in parallel while varying their individual weight to the whole grade. At the introductory stage, syntax had more weight. Similarly, at the intermediate stage, semantics had more weight than the other properties and at the advanced stage, results had more weight.

To carry out partial grading, we used software correctness levels and to enable dynamic grading, we used varying weights for properties like syntax which addressed different skills levels as we will see in this paper. We graded more than 25 different SQL questions for 309 students using different configurations of dynamic grading and we shall discuss 3 of these questions.

1.1. Advantages of skill based grading

Grading students' work while considering whether they are novice or more advanced can have several advantages. This is in comparison with grading the work of students who are struggling with the new programming language and those who are already familiar with or mastered it the same way.

1. Personalization and differentiation. Grading students work while considering whether they are novice or advanced would be considered as a personalized and differentiated way. In this manner, novice students are not penalized for not meeting a higher standard as expected of the more advanced students.

2. Student motivation and engagement. By basing the grading on the student knowledge level, their motivation to continue learning and engaging would be protected. This is because, the students are able to view learning the new programming language as something within their capability.

3. Targeted feedback. Due to the differences between novice students and advanced ones, specific feedback can be constructed that caters to the differing needs of both. Novice students may be more interested in the language syntactic elements while advanced students being more interested in semantic elements.

2. Related work

Partial grading systems have enabled awarding partial points instead of strict 0 or 1 [2, 3, 4, 5]. Some systems are even capable of generating datasets for testing [13]. The limitation of these systems is that they employ inflexible strategies to grade the queries. Our model is able to dynamically grade queries based on the gradual learning process of students. Furthermore, it has the flexibility of adding or removing properties during grading depending on the goals of the instructor.

Correctness levels have been used before to grade SQL queries. Some research has used 8 different correctness levels to partially grade queries [14]. Specifically, Dekeyser et al. [14] use syntax, schema, results, and peer review to grade the queries whereby each code feature is evaluated as either correct or not. Of these 8 levels, 5 are automatically awarded by the system while the rest require manual input. Our model automatically awards 9 different levels using syntax, semantics, and results, evaluating each feature using three categories instead of two. Our model also allows more levels by having the flexibility of adding more properties and their outcomes.

Dynamic grading has been used in other areas apart from SQL. For example, in gradual programming whereby the students learn a programming language in well-defined steps instead of learning everything at once [15].

Binary grading offers only two possible outcomes: correct and incorrect, which does not consider partial correctness or incorrectness [16]. As a result, students may not receive points for partially correct submissions. Providing detailed feedback on the areas where students' understanding is lacking becomes impossible as well. This limitation of binary grading gives rise to the need for partial grading. Our model offers more possible outcomes to enable differentiation of errors.

There are differences between novice and expert programmers. As a result, various research has been proposed which highlights these differences. It has been proposed that expert programmers memorize source code better than novice programmers [17]. In the proposal, the researchers also argued that novices tend to memorize syntactic elements while expert programmers focus on the algorithm. Expert programmers can solve a problem by abstracting the algorithm while novices cannot [18]. As the expertise increases, novice programmers gradually shift from focusing on syntactic elements to the semantic ones [19]. This may indicate that novice programmers are more focused on the language syntax while experts focus on the semantics. Differences among novices and experts have been argued in other fields as well. For example, in physics, it has been proposed that experts might use abstraction to solve a problem while novices use the problem's literal features [20]. Naturally, computer science students as programmers might show this phenomenon. Even though they may not become full experts in the programming language during the course of the class, they are able to achieve competency [21]. Due to the differences between novice and advanced students, it has been proposed to make the introductory material simple and systematically expand as students' expertise increases [22]. For example, beginning with writing syntax, comprehending templates, writing code with templates and tracing [23]. Therefore, we considered and utilized these findings to implement our grading model in such a way that it can acknowledge and accommodate these differences.

Grading students work based on unit testing has also been used in other programming languages like Java [24, 25]. Unit testing [26] involves testing the smallest unit in a program or code that can be isolated. In most cases, the testing parameters are already known by the developers and the testers before runtime. Our syntax analysis is a form of unit testing whereby a student's query's syntax is verified based on the target SQL standard and other similar syntactically correct queries. However, neither the testing parameters for our semantic nor results analysis are known up until runtime whereby the system loads the model query and the student query. Also, it is challenging to exactly anticipate the correct semantics since queries written in different ways can have the same meaning. Furthermore, the binary nature

of unit testing makes it difficult to carry out partial grading. This means that, our unit tests have additional requirements compared to ordinary unit tests.

3. Software correctness levels

Software correctness in software exercises is the degree to which code written by students satisfy the specifications given by the instructor [27]. Software correctness levels describe this degree [28]. Software correctness in SQL is tested using various techniques. The most used is executing the query and checking the results [29, 30].

There are various properties that can be evaluated to check for code correctness. Some of these properties are: syntax, semantics, results, efficiency, style and performance. During this evaluation, we classify a property into various categories that describe the correctness of the property. In this research, we call such categories ***outcomes***. Examples of outcomes are fully correct, meaning the property completely satisfies the specifications and fully incorrect for otherwise.

Using these properties and their outcomes, we can create a matrix of all the outcomes arranged in such a way that the first row describes the lowest adherence of the properties to the specifications. The last row would describe the complete adherence of these properties to the given specifications. We show such matrices as tables: Table 1, Table 2 and Table 3.

The number of properties and outcomes to use depends on the grading goals and the programming language. Database managements systems feature query optimizations therefore in SQL we would not evaluate properties like efficiency [31]. In C++, style, complexity and efficiency have been used to evaluate students' code [32, 33, 34]. In this implementation we focus on three properties: syntax, semantics, and results. for these properties, we evaluate them into three outcomes: *correct, minor incorrect*(small mistakes like misspells of 2 characters and less) and *incorrect*. The meaning behind each outcome will be discussed in the method section.

3.1. Property weight

Teaching SQL in a computer science class is done in various stages [10]. One of the most important stages is teaching proper SQL syntax. In this stage, students are introduced to SQL, how it is written, the tools used and how to run commands. Another notable stage is using already learned SQL commands to carry out simple tasks. In this stage students learn how to take simple specifications and turn them into SQL commands that accomplish the given task. Various stages of learning have various goals. Naturally, this should translate to grading goals. Our goal is to modify how grading is done based on the learning stage.

Grading modification is enabled in our model by assigning different weights for the properties. For example, At the introductory stage, grading would weigh more on syntax. During the intermediate stage, grading would weigh more on semantics, as the instructor turns their attention to carrying out given tasks. Similarly, at an advanced stage, grading would weigh more on results.

3.1.1. Introductory stage grading

At the introductory stage, the instructor would focus on correct syntax. For a student to successfully pass this stage, they would need to be able to write well-formed SQL clauses with correct keywords. For any quizzes at this stage, the instructor would classify the student queries based on how well written the syntax was. At one extreme of this classification, there would be completely incorrect syntax and at the other end fully correct syntax. The instructor would then put a threshold for passing, whereby, students below this threshold would have to practice more on syntax. A 9-magnitude model weighing more on syntax is shown in Table 1. In this model, the instructor might put the threshold for passing at correctness level 6. Students who can pass this threshold would be able to write queries that can be parsed and well executed by the database management system. Even though in this stage we focus on correct SQL syntax, the other properties, *semantics* and *results* still matter.

Table 1: Syntax has the most weight, next semantics and then results. inc. refers to incorrect.

Level	Syntax	Semantics	Results	Grade
1	incorrect	incorrect	incorrect	0
2	minor inc.	incorrect	incorrect	0.125
3	minor inc.	incorrect	minor inc.	0.25
4	minor inc.	minor inc.	correct	0.375
5	minor inc.	correct	correct	0.5
6	correct	incorrect	incorrect	0.625
7	correct	incorrect	minor inc.	0.75
8	correct	minor inc.	correct	0.875
9	correct	correct	correct	1

3.1.2. Intermediate stage grading

At an intermediate stage, the instructor would turn the students' attention to solving some tasks. At this point the students already know how to write correct SQL grammar. The instructor would teach how to turn given specifications into correct SQL queries that accomplish the task. For any given quiz at this stage, the instructor would want to check if the students are able to carry out some tasks using SQL. This would involve checking if the students' queries have the correct semantics. A 9-magnitude model focusing more on semantics is given in Table 2. In this model, the instructor might put a passing threshold at correctness level 6. This means that students with queries below this threshold might be struggling with how to turn some specifications into a semantically correct query.

3.1.3. Advanced stage grading

In the last stage, the students should already have mastered how to write correct SQL grammar. Furthermore, they should be able to turn given specifications into semantically correct SQL statements. At this point, the instructor might focus on making sure the students are able to write queries that return the expected results. This means, grading would weigh more on results. A 9-magnitude model weighing

more on results is shown in Table 3. Similarly, as the other stages, the instructor might put the threshold for passing at correctness level 6. Correct semantics imply correct results. Nevertheless, instead of focusing on semantics only, we check results because they give extra information about those queries without fully correct semantics. For example, if the reference correct query output is a subset of a student query, it means the student query may not have carried out correct filtering of results in the WHERE clause.

Table 2: Semantics has the most weight, next syntax and then results. inc. refers to incorrect.

Level	Semantics	Syntax	Results	Grade
1	incorrect	incorrect	incorrect	0
2	incorrect	minor inc.	incorrect	0.125
3	incorrect	minor inc.	minor inc.	0.25
4	incorrect	correct	incorrect	0.375
5	incorrect	correct	minor inc.	0.5
6	minor inc.	minor inc.	correct	0.625
7	minor inc.	correct	correct	0.75
8	correct	minor inc.	correct	0.875
9	correct	correct	correct	1

Table 3: Results has the most weight, next semantics and then syntax. inc. refers to incorrect.

Level	Results	Semantics	Syntax	Grade
1	incorrect	incorrect	incorrect	0
2	incorrect	incorrect	minor inc.	0.125
3	incorrect	incorrect	correct	0.25
4	minor inc.	incorrect	minor inc.	0.375
5	minor inc.	incorrect	correct	0.5
6	correct	minor inc.	minor inc.	0.625
7	correct	minor inc.	correct	0.75
8	correct	correct	minor inc.	0.875
9	correct	correct	correct	1

The instructor is free to use any threshold for passing that aligns with their goals. Notice how in all the different weights models, the students are still allowed to make minor mistakes without severe points reduction in their grades. This gives appreciation to their efforts in carrying out the tasks.

3.2. Grading

The final grade awarded to a query is calculated using the correctness level assigned as shown below:

$$g = \frac{l - min(l)}{max(l) - min(l)} \quad (1)$$

where g is the grade in the range 0 to 1, l is the correctness level, $min(l)$ is the minimum correctness level and $max(l)$ is the maximum correctness level of a specific model in use. The grade is linearly distributed along the possible correctness levels. This is an improvement to the earlier model which factored in the missing or impossible levels when calculating the grade [28].

3.3. Flexibility in adding or removing properties

Our software correctness model is made up of combinations of the chosen properties and their possible outcomes. The 9-magnitude model shown in Table 1, Table 2 and Table 3 comprises of properties *results*, *semantics* and *syntax* each with 3 possible outcomes: correct, minor incorrect and incorrect. This gives rise to 27 possible combinations of which 9 are usable in our implementation. Unusable combinations are for example, when both syntax and semantics are fully incorrect while the results, fully correct. The instructor is free to choose whichever properties or outcomes align with their teaching goals. For example, they can choose to carry out binary grading using only the results. In this case, property *results* would have 2 possible outcomes: correct and incorrect. This would translate in the correctness model having 2 levels: level 1 for an incorrect query and level 2 for a correct query. Using the same property, they might opt for a 3 magnitude correctness levels model. In this case, results would have 3 outcomes: *fully correct*, *minor incorrect* for results not filtered, i.e. correct results would be contained inside the query results and *fully incorrect*. Similarly, the instructor can choose syntax and semantics, or results and syntax with various outcomes.

As we see here, the number of properties or outcomes used to construct a correctness model for grading is not fixed. The instructor can add or remove properties and their outcomes as well. Adding properties or outcomes increases the size of the correctness model. This increases the number of correctness levels as well, which translates to higher grading sensitivity. On the other hand, using fewer properties or outcomes reduces the size of the correctness model. This reduces the number of correctness levels, which translates to lower grading sensitivity. In this paper, we demonstrate the power of a correctness model using 3 properties and 3 outcomes. Using 4 properties or outcomes and more or less is also possible in this model.

4. Grading process

Algorithm 1 shows the main algorithm for grading students queries using the correctness model discussed above. This was implemented in C++ under Linux with DuckDB [35] as the database management system.

4.1. Initialization

We begin by initializing parameters that will be used to instantiate the correctness model. This involves initializing the number of syntax, semantics and results outcomes that will be used on grading. We also initialize the 'property weight' variable. We assert that these variables are initialized to integers greater than 0 and a maximum of 3. This maximum is not fixed but can change by adding more properties or outcomes. A correctness model matrix is constructed using this initialization as shown in algorithm 2. This is a 2-dimensional array containing all the various combinations of the properties and their outcomes. We then remove unusable rows from the matrix.

Algorithm 1: Main Algorithm

```
Result: Grades
stx ← number of syntax outcomes;
sem ← number of semantics outcomes;
res ← number of results outcomes;
pw ← property weight;
// Ensure 1 ≤ stx ≤ 3, 1 ≤ sem ≤ 3, 1 ≤ res ≤ 3,
   1 ≤ pw ≤ 3
text_ed ← 3; // text edit distance
tree_ed ← 2; // tree edit distance
incorrect ← 0;
minor incorrect ← 1;
correct ← 2;
// See Algorithm 2
correctness_m ← create_matrix(stx, sem, res, pw);
// Put reference queries into data
   structure
ref ← reference queries;
// Load student queries
student ← student queries;
for n ∈ {ref} do
    n.AST ← reference query abstract syntax tree;
    // Run the query in DuckDB
    n.Output ← query result;
end
for n ∈ {student} do
    if n ≡ {} then
        // Empty
        n.PARSEABLE ← false;
        n.SYN ← incorrect;
    else
        Parse the query;
        if parseable then
            n.PARSEABLE ← true;
            n.AST ←
              student query abstract syntax tree;
            n.Output ← query result;
        end
        else
            n.PARSEABLE ← false;
            n.SYN ← incorrect;
        end
    end
end
foreach n ∈ {student} do
    // See Algorithm 3
    syntax_analysis(n, text_ed, syn);
    // See Algorithm 4, ref(1) denotes the
       first query
    results_analysis(n, ref(1), res);
    // See Algorithm 6
    semantics_analysis(n, ref, student, tree_ed, sem);
end
foreach n ∈ {student} do
    // See Algorithm 5
    get_correctness_level(n, correctness_m, stx, sem, res);

    get_grade(n);
end
```

Algorithm 2: Creating the correctness matrix

```
Result: Correctness matrix
Function create_matrix(stx, sem, res, pw):
    matrix ← vector of integers;
    if stx ≡ 1 and sem ≡ 1 and res ≡ 1 then
        // Binary grading
        Append 0,0,0 to matrix;
        Append 1,1,1 to matrix;
        return matrix;
    end
    if pw ≡ 1 then
        // Syntax most important
        for i = 0 to stx − 1 do
            for j = 0 to sem − 1 do
                for k = 0 to res − 1 do
                    Append k,j,i to matrix;
                end
            end
        end
    end
    else if pw ≡ 2 then
        // Semantics most important
        for j = 0 to stx − 1 do
            for i = 0 to sem − 1 do
                for k = 0 to res − 1 do
                    Append k,j,i to matrix;
                end
            end
        end
    end
    else if pw ≡ 3 then
        // Results most important
        for k = 0 to stx − 1 do
            for j = 0 to sem − 1 do
                for i = 0 to res − 1 do
                    Append k,j,i to matrix;
                end
            end
        end
    end
    // Remove the levels not possible from
       the matrix
    return matrix;
```

4.2. Pre-processing

Next, we carry out pre-processing of reference and student queries. For reference queries, we create abstract syntax trees for each and execute them, saving the output in a 2-dimensional array. We used the library, pg_query [36] to create abstract syntax trees. Student queries undergo the same procedure but first, we check if they are parseable. If a query is not parseable, we mark its syntax outcome as incorrect. Then, we carry out syntax, results and semantics analysis finalizing by calculating the correctness level for each query and the subsequent grade. These steps will be discussed next.

4.3. Syntax analysis

Syntax analysis is done first. Algorithm 3 shows how syntax analysis is accomplished. In this stage, all parseable queries are marked as having correct syntax. Next, we check how far the unparseable queries are from being parseable. To accomplish this, we carry out a comparison between unparseable queries with those that are parseable using text edit distance metric [37]. The parseable queries include both parseable reference and student queries. For those student queries with mistakes of 2 characters and below, we mark as having minor incorrect syntax and edit them. This enables further processing of the queries. For unparseable queries without close reference or correct student queries, we then check the SQL keywords in the query. We employ the same edit distance metric to check how much malformed the keywords are. We identify the SQL keywords using information from the parse tree.

Algorithm 3: Syntax analysis

```
Result: Syntax outcomes
Function syntax_analysis(query, text_ed, syn):
    if query.PARSEABLE ≡ true then
        query.SYN ← correct;
    end
    else
        mis ← misspelled SQL keywords characters;
        if mis < text_ed then
            if syn ≡ 3 then
                query.SYN ← minor incorrect;
                Edit query; // Correct the query
            end
            else
                query.SYN ← incorrect;
            end
        end
        else
            query.SYN ← incorrect;
        end
    end
```

4.4. Result analysis

Result analysis is done next. Algorithm 4 shows how results analysis is accomplished. In this stage, we compare the results of each student query with the results of the reference queries. If both outputs match, the student query is deemed as having correct results. For those queries with different output, we then check if the reference query output is a subset of the student query output. If so, we mark the student query as having minor incorrect results. The rest of cases are marked as having fully incorrect results.

4.5. Semantics analysis

The final analysis involves semantics. Algorithm 6 shows how semantics analysis is accomplished. In this stage we check if the student queries have the intended meaning. First, we mark as having correct semantics, those queries that had correct results in the previous step. Next, we mark unparseable queries as having fully incorrect semantics. This is because in this implementation we cannot verify the severity of the mistakes.

For the rest, we carry out parse tree edit distance comparisons using Shasha Zhang's algorithm [38]. This is accomplished by comparing parse trees of the student queries under processing and the reference queries together with the student queries with correct output. We further check the text edit distance for those queries found to have a tree edit distance of 1. If this text edit distance is 2 characters or less, the query is marked as containing minor incorrect semantics. The rest of the cases are marked as having fully incorrect semantics.

Algorithm 4: Result analysis

```
Result: Results outcomes
Function result_analysis(query, ref, res):
    if query.OUTPUT ≡ ref.OUTPUT then
        query.RES ← correct;
    end
    else if ref.OUTPUT ⊂ query.OUTPUT then
        if res ≡ 3 then
            query.RES ← minor incorrect;
        end
        else
            query.RES ← incorrect;
        end
    end
    else
        query.RES ← incorrect;
    end
```

Algorithm 5: Calculating correctness level

```
Result: Correctness level of a query
Function
 get_correctness_level(query, matrix, stx, sem, res):
    level ← 1;
    if stx ≡ 1 and sem ≡ 1 and res ≡ 1 then
        // Binary grading
        level ← query.RESULT  1;
        return level;
    end
    foreach m ∈ {matrix} do
        if m0 ≡ query.RES and m1 ≡ query.SEM and
          m2 ≡ query.SYN then
            break;
        end
        level ← level  1;
    end
    return level;
```

At this point, the student queries have been tagged with various outcomes for the properties. The outcomes of each query are compared with the outcomes in the correctness matrix. The correctness level of a query is the level with matching outcomes. The grades are calculated in the range 0 to 1, using Equation 1.

Algorithm 6: Semantics analysis

```
Result: Semantic outcomes
Function
 semantic_analysis(query, ref, student, tree_ed, sem):
    if query.RES ≡ correct then
        query.SEM ← correct;
    end
    else
        if query.PARSEABLE ≡ false then
            query.SEM ← incorrect;
        end
        else
            smallest ← 0;
            ted ← 0;
            foreach m ∈ {ref} do
                dist ←
                 query.AST and m.AST distance ;
                l_ted ←
                 query.QUERY and m.QUERY distance ;

                if dist ≤ smallest and l_ted ≤ ted then
                    smallest ← dist;
                    ted ← l_ted;
                end
            end
            foreach s ∈ {student} do
                if s.RES ≡ correct then
                    dist ←
                     query.AST and s.AST distance ;
                    l_ted ←
                     query.QUERY and s.QUERY distance;

                    if dist < smallest and l_ted ≤ ted
                     then
                        smallest ← dist;
                        ted ← l_ted;
                    end
                end
            end
            if smallest < tree_ed and ted < text_ed
             then
                if sem ≡ 3 then
                    query.SEM ← minor incorrect;
                end
                else
                    query.SEM ← incorrect;
                end
                query.RES ← correct;
            end
            else
                query.SEM ← incorrect;
            end
        end
    end
```

5. Implementation and findings

We experimented with the correctness model discussed above for an automated grading system of SQL exercises in a first year's course at a first year's BSc course, Information Modelling and Databases at Radboud University. During this study, the course had 309 students. For the year 2023, we used the model to grade 5 quizzes containing more than 25 different questions using the three different matrices. These are shown in a cumulative graph, Figure 1. From the graph, we see that the grading model can differentiate student answers among 9 different groups. We will discuss 3 results from this large sample.

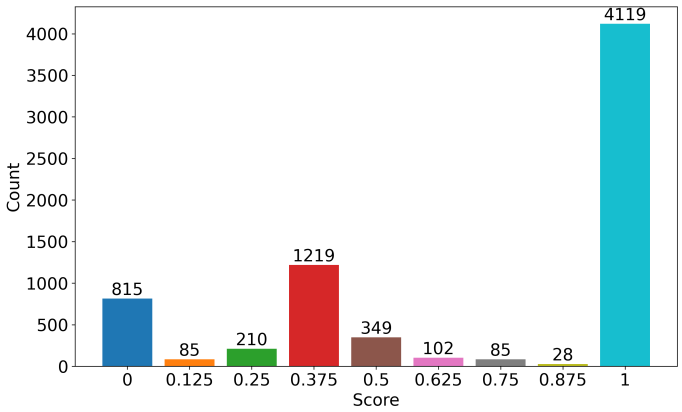


Figure 1: A bar graph showing the grades awarded to all 7012 answers across 25 different questions.

5.1. Introductory stage grading

The first question was graded as an introductory quiz. Syntax was prioritized more than the other properties. The grades awarded for this question are shown in Figure 2. The x-axis shows the grades given in the range 0 to 1. The y axis show the total number of queries given a certain grade.

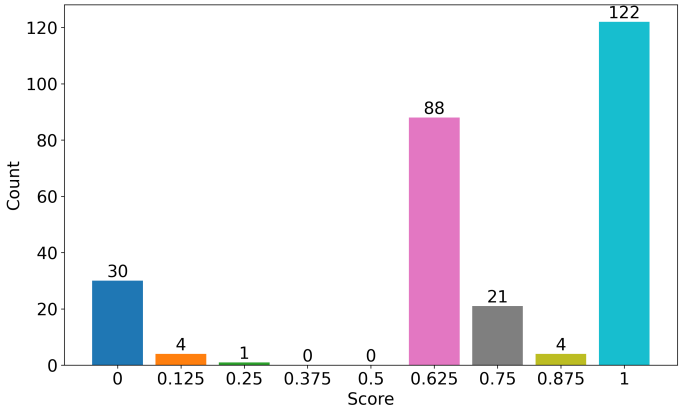


Figure 2: Grades calculated for the 1^{st} question using introductory grading.

We saw that most students were able to write correct SQL statements (122 students). This group of students could already write syntactically correct SQL that accomplished the given task. The second largest group could form correct syntax SQL queries but could not yet translate these into semantically correct queries (88 students). Taking 0.6 as the passing threshold (parseable queries), 35 students represented by 0, 0.125 and 0.25 grades failed the first question. The other 235 students passed. Grading the same question while prioritizing semantics would have produced

the results shown in Figure 3. In this grading, the number of students who passed would have reduced to 126. Lastly, grading the question while prioritizing results would have produced the results shown in Figure 4. In this case, the number of students who passed would have stayed the same as in previous case. However, the grades of those who failed would have reduced even further especially the large group of 88 answers. The reference correct query for this question is shown below.

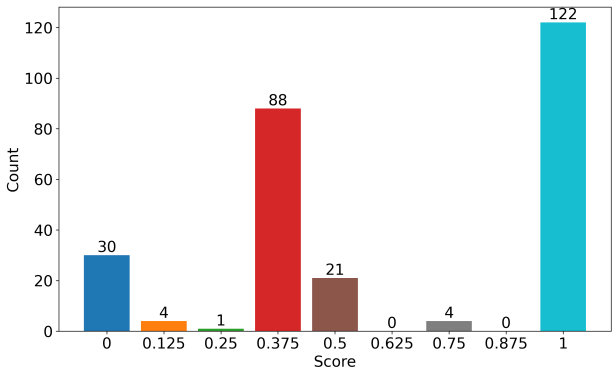


Figure 3: Grades calculated for the 1st question using intermediate grading.

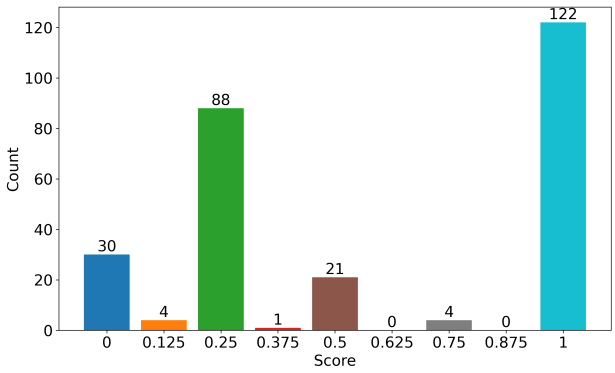


Figure 4: Grades calculated for the 1st question using advanced grading.

```
SELECT T.theme_id, T.name FROM Theme T, Teacher N
    WHERE N.name='Djoerd Hiemstra' AND
        T.teacher_id=N.teacher_id;
```

The student answers for this question fell into the following groups:

(1) 122 answers were fully correct.

```
--various query construction with the same meaning
SELECT theme_id, Theme.name FROM Teacher JOIN Theme
    USING (teacher_id) WHERE Teacher.name = 'Djoerd
    Hiemstra';
SELECT Theme.theme_id, Theme.name FROM Teacher JOIN
    Theme ON Teacher.teacher_id = Theme.teacher_id
    WHERE Teacher.name = 'Djoerd Hiemstra';
```

(2) 4 answers had minor incorrect semantics (tree edit distance of 1 and text edit distance of less than 3 from a reference query), and correct syntax and results.

```
-- mispelled the name 'Djoerd Hiemstra'
SELECT Theme.theme_id, Theme.name FROM Teacher,
    Theme  WHERE Teacher.name = 'Djoerd Hiemsta' AND
    Teacher.teacher_id = Theme.teacher_id;
```

(3) 21 answers had minor incorrect results (output of the reference correct query is a subset of the query's output), incorrect semantics and correct syntax.

```
--results were not well filtered
SELECT theme_id, Theme.name FROM Teacher, Theme
    WHERE Teacher.name = 'Djoerd Hiemstra';
```

(4) 88 answers managed only correct syntax.

```
--wrong columns selected and wrong filtering
SELECT * FROM Teacher, Theme WHERE Teacher.name =
    'Djoerd Hiemstra';
```

(5) 1 answer had minor incorrect syntax (text edit distance of less than 3 from a reference query) and results, and incorrect semantics.

```
--mispelled 'SELECT' and wrong filtering
SELECHT theme_id, Theme.name FROM Teacher, Theme
    WHERE Teacher.name = 'Djoerd Hiemstra';
```

(6) 4 answers had minor incorrect syntax, and fully incorrect results and semantics.

```
-- extra wrong character ';', wrong columsn selected
    and wrong filtering
SELECT * FROM Teacher A, Theme T; WHERE T.name =
    'Djoerd Hiemstra';
```

(7) 30 answers had fully incorrect properties.

```
SELECT name AND theme_id FROM Theme WHERE name =
    Djoerd Hiemstra
```

From the three graphs (Figure 2, Figure 3, and Figure 4), we see that the grouping stays the same. What changes is the grades awarded to the various groups seen. We also see that the grading progressed from lenient to stringent as we progress from grading using syntax, to semantics and finally results. For this reason, we focus on syntax to carry out introductory stage grading whereby we do not penalize novice students for not meeting higher standard as expected of the more advanced students. We strive to make an impression to the novice students that the new language presented is within their learning capability.

5.2. Intermediate stage grading

As the course progressed the next questions were graded as intermediate quizzes. In this subsection, we give an example of a question which was graded in this intermediate stage whereby, semantics were prioritized more than the other properties. The grades awarded for this question are shown in Figure 5. In the grades evaluated, we saw that almost the whole class (230/262) was able to write semantically fully correct queries. 18 students (0, 0.125 and 0.375 grades) got less than a half point therefore failed this exercise. The reference correct query for this question is shown below.

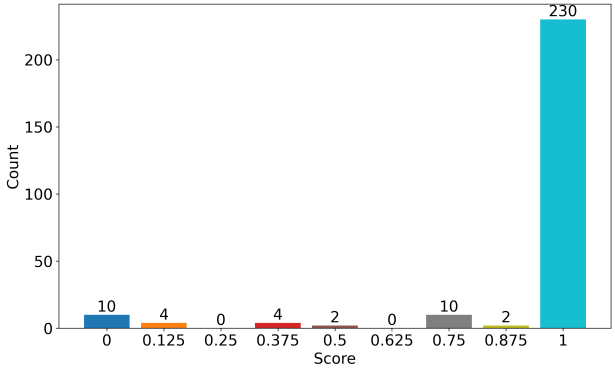


Figure 5: A bar graph showing the first question of the intermediate quiz. This was graded with more weight being on semantics.

```
SELECT DISTINCT album_title FROM Album WHERE
    type='compilation';
```

The student answers for this question fell into the following groups:

(1) 230 answers were fully correct.

```
--different phrasing but still correct
SELECT Album.album_title FROM Album WHERE
    Album.type='compilation';
```

(2) 2 answers answers had minor incorrect syntax, and correct semantics and results.

```
--extra ';' which made syntax minor incorrect
SELECT album_title; FROM Album; WHERE type =
    'compilation';
```

(3) 10 answers had minor incorrect semantics, and correct results and syntax.

```
-- missing enclosing ' for the word compilation
SELECT album_title FROM Album WHERE type =
    compilation
```

(4) 2 answers had minor incorrect results, fully incorrect semantics and correct syntax.

```
--selected more data than required. Correct results
    were a subset.
select album_title, type from album where type =
    'compilation'
```

(5) 4 answers had fully incorrect results and semantics, and fully correct syntax.

```
-- the query could be parsed even though the where
    clause was incorrect.
Select album_title from Album where album_name
```

(6) 4 answers had fully incorrect results and semantics, and minor incorrect syntax.

```
--had an extra '*' and the wrong column name
    'tracck_title'.
SELECT * track_title FROM Album WHERE
    type='compilation';
```

(7) 10 answers had fully incorrect properties.

```
sigma type='compilation'(Album) pi album_title
```

5.3. Advanced stage grading

As the course progressed to the final lessons, the questions given were graded as advanced quizzes. In this subsection, we give an example of a question which was graded in this advanced stage whereby, results were prioritized more than the other properties. The grades awarded for this question are shown in Figure 6. In these results, we saw that many students (110/164) managed to write fully correct queries. The second largest group (29/164) students wrote a fully incorrect query even though this group was small compared to those who wrote fully correct queries. The rest of the small groups were distributed among various scores. The reference correct query for this question is shown below.

```
SELECT name, hire_date FROM Employee WHERE salary
    BETWEEN 2500.00 AND 3500.00;
```

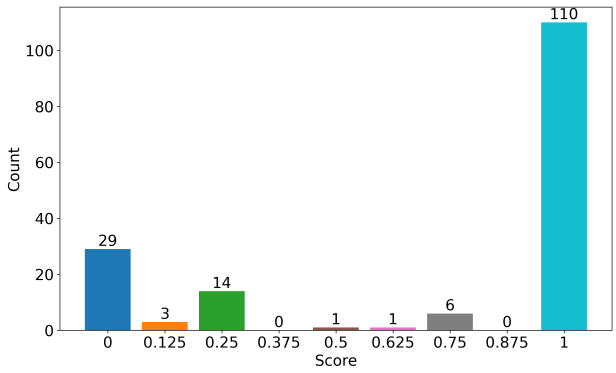


Figure 6: A bar graph showing one of the last questions of the advanced quiz. This was graded with more weight being on results.

The student answers for this question fell into the following groups:

(1) 110 answers were fully correct.

```
--different phrasing but still correct
SELECT name, hire_date  FROM Employee  WHERE salary
    >=2500 AND salary <= 3500;
```

(2) 6 answers had fully correct results and syntax but minor incorrect semantics.

```
-- had an extra 's' at the end of the string
    'Employee'
SELECT name, hire_date FROM Employees WHERE salary
    BETWEEN 2500 AND 3500;
```

(3) 1 answer had fully correct results but minor incorrect syntax and semantics.

```
-- incomplete where clause, missing second string,
    'salary' after AND
SELECT name , hire_date FROM Employee WHERE salary
    >= 2500 AND <= 3500;
```

This single query was incorrectly graded due to an error in the implementation which we rectified afterwards.

(4) 1 answer had fully correct syntax but minor incorrect results and fully incorrect semantics.

```
-- added another table at the FROM clause
SELECT name, hire_date  FROM Unit, Employee  WHERE
    salary > 2500  AND salary < 3500;
```

(5) 14 answers had fully correct syntax but fully incorrect results and semantics.

```
-- missing 'hire_date' column in SELECT clause
SELECT name FROM Employee WHERE salary>2500 AND
    salary<3500
```

(6) 3 answers had minor incorrect syntax and fully incorrect results and semantics.

```
--missing 'hire_data' in SELECT clause and 'salary'
    in WHERE clause
SELECT name  FROM Employee  WHERE salary IS BETWEEN
    2500 AND 3500
```

(7) 29 answers had fully incorrect semantics, results and syntax.

```
SELECT name, hire_date  FROM Employee, Unit   WHERE
    2500 < salary < 3500
```

5.4. Comparison with manual grading

To validate the capabilities of our implementation, we carried a comparison between the grades awarded for the end exam by the instructor, and those calculated by our automated grading system. The instructor evaluated the students not as novices but as learners already familiarized with SQL concepts and knowledge. Our grading system evaluated the queries as advanced quizzes to match the instructor's grading strategy using the model shown in Table 3. The correct answers to the questions are shown below.

```
1. SELECT A.name, COUNT(DISTINCT T.track_id) AS
    number FROM Artist A, Performs P, Track T WHERE
    A.artist_id = P.artist_id  AND P.track_id =
    T.track_id GROUP BY A.name HAVING COUNT(*) >= 3
    ORDER BY number DESC;
2. SELECT title, SUM(duration) FROM Album A, Track T
    WHERE A.album_id = T.album_id GROUP BY title;
3. SELECT song FROM Track T WHERE NOT EXISTS( SELECT
    * FROM Performs P WHERE P.track_id = T.track_id
    AND role='vocals') AND EXISTS( SELECT * FROM
    Performs P WHERE P.track_id = T.track_id);
4. SELECT song FROM Album A, Track T WHERE
    A.album_id=T.album_id AND A.title='Nevermind';
5. SELECT album_id  FROM Release WHERE
    country='Belgium';
6. SELECT DISTINCT artist_id FROM Performs WHERE
    role = 'guitarist';
7. SELECT name FROM Artist WHERE death_date IS NULL;
8. WITH RECURSIVE Subordinates(employee_nr) AS
    (SELECT 'U000001' UNION SELECT
    Employee.employee_nr FROM Employee, Subordinates
    WHERE Employee.manager_nr =
    Subordinates.employee_nr ) SELECT * FROM
    Subordinates;
```

In the comparison, we checked the difference to the grades given for each question using the equation shown below:

$$difference = our_grade - instructor_grade \quad (2)$$

The frequency of differences seen are summarized in Table 4. In the table, score diff. is the difference between the grade given by the automated grading system and by the instructor. Q1 to Q8 are the 8 questions. Each cell in the questions' columns contain the total number of queries with a given grade difference in that question. The total column show the total number of queries for a given difference in all the questions. These are further illustrated using violin plots in Figure 7. In the figure, Q1 to Q8 are the eight questions. Each question has its own violin plot. The score difference is the difference between the grade given by the automated grading system and by the instructor.

Table 4: The frequencies of differences. Q1 to Q8 are the 8 questions. Diff. is the difference value.

Score Diff.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
-1.0	1	0	1	1	0	0	0	1	4
-0.875	0	0	0	0	1	0	0	0	1
-0.75	9	21	0	1	0	1	0	0	23
-0.6	1	0	0	0	0	0	0	0	1
-0.55	0	2	2	41	43	31	0	2	121
-0.5	3	0	0	0	0	0	0	0	3
-0.35	0	0	0	0	0	1	0	0	1
-0.25	47	0	0	0	1	0	0	0	48
-0.05	0	0	7	0	0	0	0	0	7
0.0	63	123	98	180	171	146	202	110	1093
0.125	14	1	0	1	0	0	3	8	27
0.2	0	0	0	1	0	0	0	0	1
0.25	108	103	96	22	30	50	26	115	550
0.375	0	0	1	0	0	0	0	0	1
0.5	1	0	48	3	2	23	1	3	81
0.625	0	0	0	1	0	0	0	0	1
0.75	0	2	0	2	6	2	22	0	34
1.0	0	0	0	1	0	0	0	0	1

We calculated the mean absolute difference (MAD) and the root mean square error (RMSE) between the two types of grading which was 0.16 and 0.07 respectively using Equation 3 and Equation 4.

$$MAD = \frac{1}{n}\Sigma_{i1}^{n}|x_i - y_i| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i1}^{n}(x_i - y_i)^2} \quad (4)$$

Whereby n is the total number of differences, x is a grade given by our system and y is a grade given by the instructor. This difference between the two systems comes from the following main observations which are summarized below:

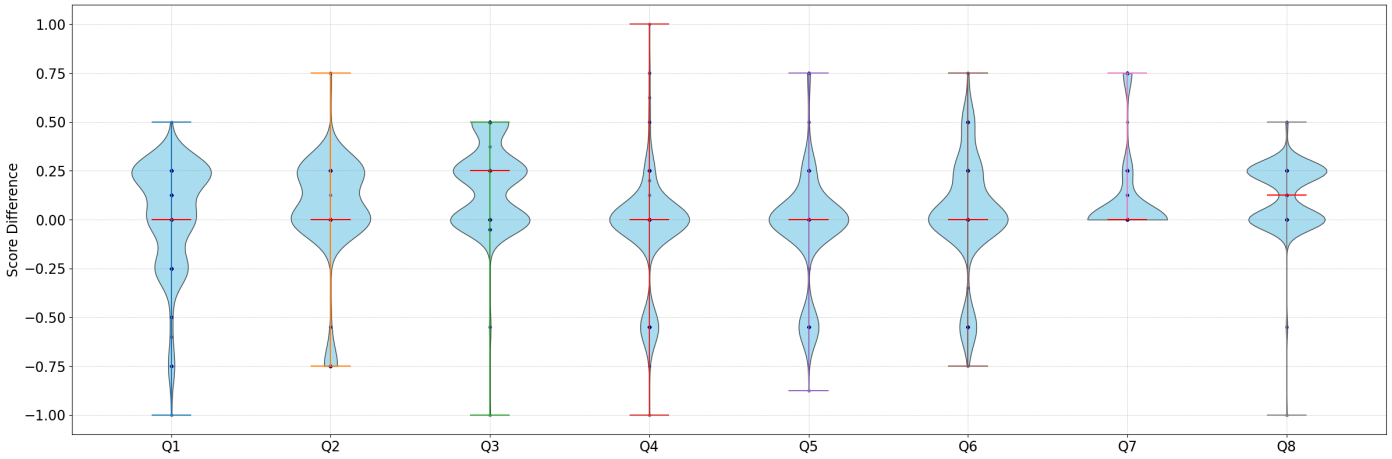


Figure 7: Grades differences between the instructor and our system. Q1 to Q8 are the 8 questions.

1. 1093 queries received the same score in both cases. These were fully correct queries and fully incorrect queries.

2. 550 queries received a 0.25 score higher in our automated grading system. These were incorrect queries that were at least parseable. Our grading acknowledged that even though these students did not learn correct SQL semantics, they were able to form correct SQL syntax.

3. 121 queries received a 0.55 score higher from the instructor. These were correct queries containing unexpected special characters, for example:

```
-- has an extra non-SQL characters Â‘
 SELECT T.song FROM Track T, Album A WHERE
     T.album_id = A.album_id AND A.title = Â
     ‘NevermindÂ‘ ;
```

```
--used non-SQL character ‘ instead of '
SELECT album_id FROM Release WHERE country
    = ‘Belgium‘;
```

```
--used non-SQL character ‘ instead of '
SELECT DISTINCT artist_id FROM Performs
    WHERE role = ‘guitarist‘;
```

   The instructor deducted a 0.2 from the score while our system struggled to correctly evaluate such queries.

4. 81 queries received a 0.5 score higher in our system. These were queries which were semantically incorrect with correct results as a subset of the queries' results.

5. 48 queries received a 0.25 score higher from the instructor. Most of these differences were from the first question. The instructor gave a 0.5 to those answers that demonstrated that the student clearly understood the semantics of the question but could not fully translate it to the complete query. For example:

```
SELECT name, COUNT(*) as count FROM Artist,
    Performs, Track WHERE Artist.artist_id
    = Performs.artist_id AND
    Performs.track_id = Track.track_id
    GROUP BY name HAVING count >= 3 ORDER
    BY count DESC;
```

   Our system failed to evaluate this and gave a 0.25.

6. 34 answers received a 0.75 score higher in our system. These were queries whose semantics were minor incorrect. For example, 17 queries contained:

```
SELECT name FROM Artist WHERE death_date =
    NULL;
```

   This query used a comparing operator for NULL instead of using IS NULL. The instructor observed that they were strict in this case.

7. 23 queries received a 0.75 higher score from the instructor. These were queries which were graded as fully correct by the instructor . Our system evaluated them as incorrect. For example:

```
SELECT A.title, SUM(T.duration) FROM Album
    A JOIN Track T ON A.album_id =
    T.album_id GROUP BY A.album_id;
```

   The query did not group by 'title' therefore it was considered as incorrect by our system.

As we seen in Figure 7, most differences between instructor's grading are within 0.25 score (horizontal bulges). We have highlighted the major differences seen above. In some cases, the instructor was strict as seen from the 34 queries in item 6 and item 4. In other cases, the instructor was lenient in grading as seen in queries in item 5 and item 7. In comparison, our grading was consistent across different queries. This consistency aided our system to allocate fair grades with some exceptions. These are cases whereby the system failed to capture special unexpected non-SQL characters as seen in item 3.

At the end of the experiment, we administered a questionnaire to the students. We let them know that the questionnaire was useful in that, we would use their answers

to improve the grading system. The students based their answers on the grades received since we carried out the grading using the system and sent them their results. This included some feedback text. For example, *SYNTAX: Correct! Well done. RESULTS: The output of the query is not correct. SEMANTICS: The query does not have correct semantics. This quiz was graded as an advanced quiz. We placed more weight on results, then semantics and finally syntax.* A group of 46 students present at the time participated in the questionnaire, in which we enquired about their opinion on the automatically given grades. For example, we asked them if they considered the grades received as fair or not. This was to verify that indeed the students found the grades given acceptable. The results are shown in Figure 8. We saw that 74% said the grades received were fair while the rest 26% considered the grades not completely fair. Some students observed our system being not completely fair in those cases our implementation failed to catch minor incorrect syntax, for example for those unparseable queries containing random characters that proved difficult to automatically identify and repair. We plan to work on this in our next implementation.

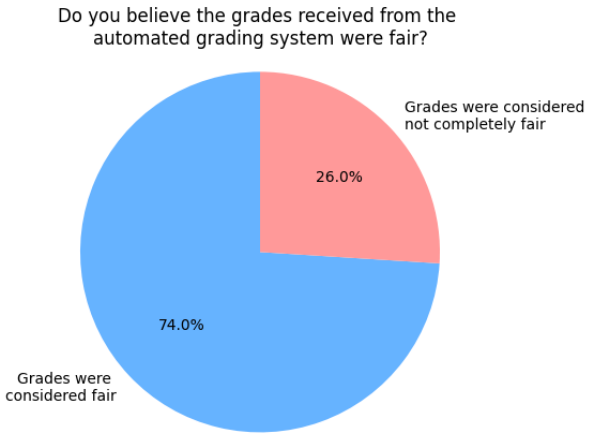


Figure 8: Survey about students' experience.

5.5. Configuration variables

In our method, there are 6 configuration variables that affect the grading (number of syntax outcomes, number of semantics outcomes, number of results outcomes, property weight, text edit distance threshold and tree edit distance threshold). Changing these variables alters the grading in different ways. The first 3 variables affect the number of outcomes for the different properties. The fourth variable affects which property is focused more when grading. The fifth variable is the threshold value for syntax analysis edit distance calculation. Lastly, the sixth variable is the threshold for the semantics tree edit distance calculation. All these variables can be adjusted to change how grading is done. The threshold parameter for edit distances for syntax and semantics in this experiment were set to 2 and 1 respectively. In our next implementation we plan to study in detail how changing these variables will affect the grades awarded.

6. Discussion

During the introductory period of our course, one of the main goals was for the students to learn the correct SQL syntax and grammar. It was crucial for the students to be able to form SQL queries that can be parsed. In our first quiz, we tested whether the students had indeed understood how syntactically correct SQL queries are written. We graded the quiz focusing mostly on syntax with semantics and results being secondary. This grading model is shown in Table 1. Our results from Figure 2 showed that this goal was accomplished as most students (235/270) were able to form syntactically correct SQL queries. These students scored more than 0.6. Furthermore, a large number (122/270) were able to write semantically correct queries. This meant that the teaching could proceed onto the next stage, which is using SQL to accomplish the given tasks.

In the next stage of learning, the students could already write syntactically correct SQL queries. This was the intermediary stage, whereby the goal was using SQL to carry out the given tasks. The quizzes in this stage were graded focusing mostly on semantics while syntax and results were secondary. This grading model is shown in Table 2. In one of the questions, shown in Figure 5, most of the students (230/262) could translate SQL into statements that could accomplish the given task.

We graded the last quizzes for the course as advanced quizzes. At this stage the students could write the correct SQL grammar and write semantically correct queries. Therefore, when grading, we focused mostly on results while semantics and syntax were secondary. Many students were able to write queries with the correct results (110/164 students). The grading model used is shown in Table 3.

7. Conclusion

In this paper, we have demonstrated a model for generating discrete levels that enable effective partial grading of SQL queries. We call these discrete levels, software correctness levels that translate into partial grades. We have also demonstrated how we can be able to offer personalized and differentiated type of grading depending on the knowledge level of the students. This is in contrast to the previous research that utilize inflexible methods. we divided the learning process into 3 stages and applied specific types of grading to each stage. These stages were introductory, intermediate, and advanced. At each stage of our course, we were able to focus the grading on a specific goal which depended on the skills of the students. This is dynamic grading. Dynamic grading was enabled by weighing some properties like syntax more than others at specific points of the teaching process. Partial grading was enabled by individually evaluating different properties in such a way that their outcomes are categorized, for example in categories like correct, minor incorrect and fully incorrect. In our current implementation, we worked with the properties syntax, semantics, and results, and we generated corresponding correctness levels for those properties. In partial grading we utilized software correctness levels while in dynamic grading, we considered students' skills level. This helped

us to effectively provide partial grades that were dependent upon the skill or knowledge level of students at different stages of their learning process.

We observed that our grading model had some differences in the grades calculated from those awarded by the instructor. We evaluated these differences from SQL questions from an exam and observed a mean average difference of 0.16. We also administered a questionnaire to 46 students whereby 74% observed that our implementation of the model discussed was able to calculate grades they considered fair. Nevertheless, there were limitations which we plan to address in both the grading model and its implementation. For the implementation, we plan on improving on: handling unexpected special non-SQL characters, checking query results against more than 1 data schema, extending the supported queries to include, *Create*, *Delete*, *Insert* and *Update* and a clear separation between syntax and semantics analysis. The later limitation resulted in missing groups for example, [syntax: minor incorrect, result: correct]. For the model, we plan on adding another property or outcome for more sensitivity and to evaluate the optimum configuration for the minimum difference between the grades awarded by the system and the instructor. More thorough work is needed to present the full merits of the discussed model.

References

[1] T. Seifert, "Understanding student motivation", *Educational Research*, vol. 46, no. 2, pp. 137–149, 2004, doi:10.1080/0013188042000222421.

[2] B. Chandra, A. Banerjee, U. Hazra, M. Joseph, S. Sudarshan, "Automated grading of SQL queries", "2019 IEEE 35th International Conference on Data Engineering (ICDE)", pp. 1630–1633, 2019, doi:10.1109/ICDE.2019.00159.

[3] G. Dambić, M. Fabijanić, A. L. Ćošković, "Automatic, configurable and partial assessment of student SQL queries with joins and groupings", "2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)", pp. 837–842, 2021, doi:10.23919/MIPRO52101.2021.9596680.

[4] M. Fabijanić, G. Dambić, J. Sasunić, "Automatic, configurable, and partial assessment of student SQL queries with subqueries", "2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)", pp. 542–547, 2022, doi:10.23919/MIPRO55190.2022.9803559.

[5] J. Kjerstad, "Automatic evaluation and grading of SQL queries using relational algebra trees", Master's thesis, Norwegian University of Science and Technology, 2020.

[6] T. J. McGill, S. E. Volet, "A conceptual framework for analyzing students' knowledge of programming", *Journal of Research on Computing in Education*, vol. 29, no. 3, pp. 276–297, 1997, doi:10.1080/08886504.1997.10782199.

[7] B. Shneiderman, R. Mayer, "Syntactic/semantic interactions in programmer behavior: A model and experimental results", *International Journal of Parallel Programming*, vol. 8, pp. 219–238, 1979, doi:10.1007/BF00977789.

[8] B. Shneiderman, "Teaching programming: A spiral approach to syntax and semantics", *Computers & Education*, vol. 1, no. 4, pp. 193–197, 1977.

[9] A. Stefik, S. Siebert, "An empirical investigation into programming language syntax", *ACM Transactions on Computing Education (TOCE)*, vol. 13, no. 4, pp. 1–40, 2013.

[10] K. Renaud, J. van Biljon, "Teaching sql — which pedagogical horse for this course?", H. Williams, L. MacKinnon, eds., "Key Technologies for Data Management", pp. 244–256, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[11] P. Garner, J. A. Mariani, "Learning sql in steps", *Journal on Systemics, Cybernetics and Informatics*, vol. 13, pp. 19–24, 2015.

[12] H. Al Shauily, K. Renaud, "A framework for sql learning: linking learning taxonomy, cognitive model and cross cutting factors", *International Journal of Computer and Systems Engineering*, vol. 10, no. 9, pp. 3105–3111, 2016.

[13] A. Bhangdiya, B. Chandra, B. Kar, B. Radhakrishnan, K. V. M. Reddy, S. Shah, S. Sudarshan, "The XDa-TA system for automated grading of SQL query assignments", *2015 IEEE 31st International Conference on Data Engineering*, pp. 1468–1471, 2015.

[14] S. Dekeyser, M. de Raadt, T. Y. Lee, "Computer assisted assessment of SQL query skills", "Proceedings of the Eighteenth Conference on Australasian Database - Volume 63", ADC '07, p. 53–62, Australian Computer Society, Inc., AUS, 2007.

[15] M. Gilsing, J. Pelay, F. Hermans, "Design, implementation and evaluation of the hedy programming language", *Journal of Computer Languages*, vol. 73, p. 101158, 2022, doi:https://doi.org/10.1016/j.cola.2022.101158.

[16] J. C. Prior, R. Lister, "The backwash effect on SQL skills grading", "Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education", ITiCSE '04, p. 32–36, Association for Computing Machinery, New York, NY, USA, 2004, doi:10.1145/1007996.1008008.

[17] M. Kramer, M. Barkmin, D. Tobinski, T. Brinda, "Understanding the differences between novice and expert programmers in memorizing source code", A. Tatnall, M. Webb, eds., "Tomorrow's Learning: Involving Everyone. Learning with and about Technologies and Computing", pp. 630–639, Springer International Publishing, Cham, 2017.

[18] M. Weiser, J. Shertz, "Programming problem representation in novice and expert programmers", *International Journal of Man-Machine Studies*, vol. 19, no. 4, pp. 391–398, 1983, doi:https://doi.org/10.1016/S0020-7373(83)80061-3.

[19] C. M. Zeitz, "Expert-novice differences in memory, abstraction, and reasoning in the domain of literature", *Cognition and Instruction*, vol. 12, no. 4, pp. 277–312, 1994.

[20] M. T. Chi, P. J. Feltovich, R. Glaser, "Categorization and representation of physics problems by experts and novices", *Cognitive Science*, vol. 5, no. 2, pp. 121–152, 1981.

[21] L. E. Winslow, "Programming pedagogy—a psychological overview", *SIGCSE Bulletin*, vol. 28, no. 3, p. 17–22, 1996, doi:10.1145/234867.234872.

[22] A. Robins, J. Rountree, N. Rountree, "Learning and teaching programming: A review and discussion", *Computer science education*, vol. 13, no. 2, pp. 137–172, 2003.

[23] B. Xie, D. Loksa, G. L. Nelson, M. J. Davidson, D. Dong, H. Kwik, A. H. Tan, L. Hwa, M. Li, A. J. Ko, "A theory of instruction for introductory programming skills", *Computer Science Education*, vol. 29, no. 2-3, pp. 205–253, 2019, doi:10.1080/08993408.2019.1565235.

[24] C. Wilcox, "Testing strategies for the automated grading of student programs", "Proceedings of the 47th ACM Technical Symposium on Computing Science Education", SIGCSE '16, p. 437–442, Association for Computing Machinery, New York, NY, USA, 2016, doi:10.1145/2839509.2844616.

[25] C. Benac Earle, L.-r. Fredlund, J. Hughes, "Automatic grading of programming exercises using property-based testing", "Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education", ITiCSE '16, p. 47–52, Association for Computing Machinery, New York, NY, USA, 2016, doi:10.1145/2899415.2899443.

[26] P. Runeson, "A survey of unit testing practices", *IEEE Software*, vol. 23, 2006, doi:10.1109/MS.2006.91.

[27] B. Wanjiru, P. v. Bommel, D. Hiemstra, "Towards a generic model for classifying software into correctness levels and its application to SQL", "2023 IEEE/ACM 5th International Workshop on Software Engineering Education for the Next Generation (SEENG)", pp. 37–40, 2023, doi:10.1109/SEENG59157.2023.00012.

[28] B. Wanjiru, P. v. Bommel, D. Hiemstra, "Sensitivity of automated SQL grading in computer science courses", "Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24)", 2024.

[29] C. Kleiner, C. Tebbe, F. Heine, "Automated grading and tutoring of sql statements to improve student learning", "Proceedings of the 13th Koli Calling International Conference on Computing Education Research", Koli Calling '13, p. 161–168, Association for Computing Machinery, New York, NY, USA, 2013, doi:10.1145/2526968.2526986.

[30] B. Chandra, B. Chawda, B. Kar, K. V. M. Reddy, S. Shah, S. Sudarshan, "Data generation for testing and grading sql queries", *The VLDB Journal*, vol. 24, no. 6, p. 731–755, 2015, doi:10.1007/s00778-015-0395-0.

[31] S. Chaudhuri, "An overview of query optimization in relational systems", "Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems", PODS '98, p. 34–43, Association for Computing Machinery, New York, NY, USA, 1998, doi:10.1145/275487.275492.

[32] K. Ala-Mutka, T. Uimonen, H.-M. Järvinen, "Supporting students in C++ programming courses with automatic program style assessment", *JITE*, vol. 3, pp. 245–262, 2004, doi:10.28945/300.

[33] F. G. Wilkie, B. Hylands, "Measuring complexity in C++ application software", *Software: Practice and Experience*, vol. 28, 1998.

[34] N. R. Tallent, J. M. Mellor-Crummey, "Effective performance measurement and analysis of multithreaded applications", PPoPP '09, p. 229–240, Association for Computing Machinery, New York, NY, USA, 2009, doi:10.1145/1504176.1504210.

[35] M. Raasveldt, H. Mühleisen, "Duckdb: an embeddable analytical database", "Proceedings of the 2019 International Conference on Management of Data", SIGMOD '19, p. 1981–1984, Association for Computing Machinery, New York, NY, USA, 2019, doi:10.1145/3299869.3320212.

[36] T. pganalyze Developer Team, "libpg_query", 2023, version 15-4.2.1.

[37] L. Yujian, L. Bo, "A normalized levenshtein distance metric", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007, doi:10.1109/TPAMI.2007.1078.

[38] K. Zhang, D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems", *SIAM J. Comput.*, vol. 18, pp. 1245–1262, 1989, doi:10.1137/0218082.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Received: 14 May, 2024, Revised: 18 July, 2024, Accepted: 19 July, 2024, Online: 01 August, 2024
DOI: https://dx.doi.org/10.55708/js0308002

MCNN+: Gemstone Image Classification Algorithm with Deep Multi-feature Fusion CNNs

Haoyuan Huang[1], **Rongcheng Cui**[*,2]

[1]College of Jewelry, Shanghai Jian Qiao University, Shanghai, 201306, China
[2]School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, 201306, China
[*]Corresponding author: Rongcheng Cui, cuirongcheng98@gmail.com

ABSTRACT: Accurate gemstone classification is critical to the gemstone and jewelry industry, and the good performance of convolutional neural networks in image processing has received wide attention in recent years. In order to better extract image content information and improve image classification accuracy, a CNNs gemstone image classification algorithm based on deep multi-feature fusion is proposed. The algorithm effectively deeply integrates a variety of features of the image, namely the main color features extracted by the k-means++ clustering algorithm and the spatial position features extracted by the denoising convolutional neural network. Experimental results show that the proposed method provides competitive results in gemstone image classification, and the classification accuracy is nearly 9% higher than that of CNN. By deeply integrating multiple features of the image, the algorithm can provide more comprehensive and significant useful information for subsequent image processing.

KEYWORDS: CNN, Multi-features, Image Classification, Gemstone, Data Fusion

1. Introduction

Gemology research [1], mainly including gemstone classification and identification of two areas, gemstone classification is the core of Gemology research is also the premise of gemstone identification, in the traditional classification of gemstone is mainly to the naked eye and gemstone microscope as the main tool, but with the progress of modern science and technology, synthetic gemstone technology is also constantly evolving, so that some of the optimized treatment of gemstone features and natural gemstone the difference is decreasing. Although complex instruments with strong spectral, fluorescence, or chemical analysis capabilities are increasingly being introduced into Gemology laboratories [2], identification is still difficult and time-consuming, and not all laboratories can specialize in precision instruments; Therefore, automatic technical recognition based only on images is attractive.

Computers and algorithms have come a long way in recent years, and image processing and computer vision tasks are common in many areas such as medical imaging, manufacturing, and security [3]. Image classification as one of the key technologies has also made great progress, mainly traditional methods and image classification methods based on deep learning. Traditional classification methods such as random forests, decision trees, and support vector machines all have a good classification effect on natural images. With the advent of big data and the rapid development of artificial intelligence, classification methods based on deep learning have become a hot topic in image classification research, and are applied to remote sensing images [4], medical images [5], and spectral images [6] and other fields. Although computer vision systems have widespread applications in many fields, there is only one study on the automatic recognition of gemstone images [7]. To date, to the author's knowledge, there have been reports. Unseen ruby, sapphire, and jadeite images are classified using artificial neural networks based on tonal channels in the HSV color space with 75-100% accuracy per class. It's important to note that rubies, sapphires, and emeralds are very unique in color and therefore relatively easy to distinguish, much easier than similarly colored gemstone such as topaz and aquamarine. Other computer vision studies in the context of Gemology focus on gemstone evaluation [8, 9, 10] and identification [11, 12]. The [13] is the first study to compare a computer vision-based approach with the image-based classification performance of trained Gemology on up to 68 types of gemstones. However, the method adopted is cumbersome and the classification accuracy needs to be improved.

In this paper, we propose a CNNs image classification algorithm based on deep multi-feature fusion for automatic image-based classification of gemstones. The work is first described in section 2 . Section 3 introduces the structure of the model we propose. The image dataset, experimental environment, and experimental results and discussion are detailed in section 4. Finally, further conclusions and ideas are presented in section 5.

2. Related Work

Accurate gemstone classification is critical to the gemstone and jewelry industry, as identification is an important first

step in evaluating any gemstone [14]. Currently, the identity of gemstone is determined by combining visual observation and spectral chemical analysis [15]. By carefully viewing the gemstone with the naked eye and a magnifying glass, Gemology can detect visual characteristics such as color, transparency, luster, fracture, cleavage, inclusions, polychromaticity, phenomena, and birefringence to facilitate the separation of the gemstone [16]. With the advent of new synthetic gemstone and treatment techniques, complex instruments with powerful spectral, fluorescence, or chemical analysis capabilities are increasingly being introduced into Gemology laboratories [15]. Such instruments include infrared spectrometers [16], Raman and luminescent spectrometers [1, 17, 18], ultraviolet-visible spectrometers [19], cathodoluminescence [20], etc.

In the geological sciences, computer vision algorithms have been developed for the study of mineral particles and rocks [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. In [21], the author segmented microscopic flakes using edge detection and achieved up to 93.53% test accuracy when classifying 10 different minerals using artificial neural networks trained on extracted color and texture features. The accuracy of the report may be exaggerated because the same Biotite examples are used for training and testing. In [22], the author developed an artificial neural network that used the red-green-blue (RGB) values of the pixels in the thin slices as input to separate five minerals with 89.53% accuracy. In [23], the author segmented flakes using incremental clustering and mineral classification using cascading methods. Artificial neural networks were first used to distinguish 23 minerals and glass based on pixel color, and only minerals that appeared similar in planar polarized and cross-polarized light were passed to a second artificial neural network for simultaneous color and texture analysis. This results in an overall accuracy of 93.81%. In [24], the author demonstrated that simple machine learning algorithms—K-Nearest Neighbor and Decision Tree—were able to classify minerals in microscopic sheets with high average accuracy of 94.11-97.71% using two datasets for four and seventeen mineral types based on color and texture. In [25], microscopic images containing eight mineral types and backgrounds were segmented by simple linear iterative clustering and classified based on RGB, hue-saturation-value (HSV) color features , and CIELAB spaces using three machine learning algorithms, K-Nearest Neighbour, Random Forest, and Decision Tree. The random forest algorithm yields the highest accuracy of 82%. In [26], the author using inception-v3 features extracted from microscopic images, the classification of four minerals was studied using six different algorithms: logistic regression, support vector machine, random forest, K-nearest neighbor, multilayer perceptron, and naïve Bayes. The support vector machine is identified as the single algorithm that produces the highest accuracy (90.6%). The Stacking Support Vector Machine, Logistic Regression, and Multilayer Perceptron models further improved accuracy by 0.3%.

3. Model

In order to better read the image content information, the overall framework of the CNN network model of deep multi-feature fusion is shown in Figure 1.

The input image is first convoluted to obtain the denoising depth convolution feature, referred to as f_c. The position shape relationship information of the denoising space after the multilayer convolutional neural network. The convolutional layer contains multiple convolutional nuclei, each of which is capable of extracting shape-dependent features, and each neuron in the convolutional layer is connected to multiple neurons adjacent to the location of the previous layer, also known as the 'receptive field' [31], thus relying on the network to learn the contextual invariant features [32]: Shape and spatial position feature information, which is particularly useful for image classification. The main color features f_l of the input image are then extracted, using the k-means++ clustering algorithm. Then, cascading f_c and f_l, construct deep-in-depth features f_m.

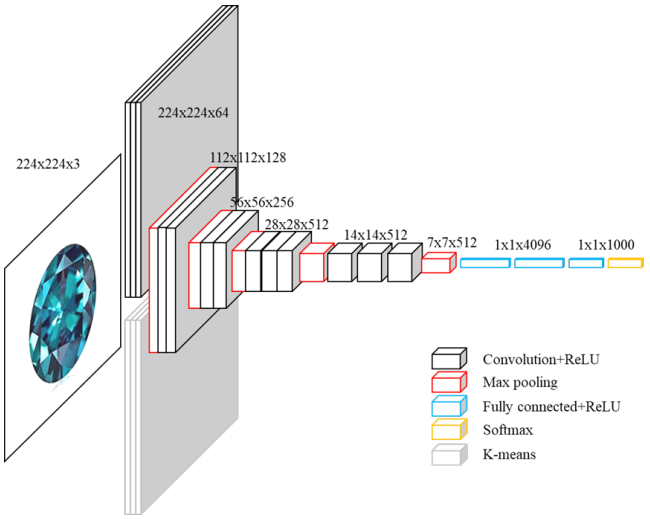


Figure 1: General framework of CNNs network model for deep multi-feature fusion

There are a total of four convolutional layers in the model frame. The first convolutional layer is called the data feature fusion layer, which fully fuses the denoising depth convolution and the main color feature. This is followed by a pooling layer whose role is to reduce network parameters and speed up fusion. The second convolutional layer is called the deep feature fusion layer, which undergoes further feature fusion, followed by a pooling layer. The third convolutional layer is called the feature abstraction representation layer, and the size of the convolutional kernel in this layer is changed from 5×5 in the first two convolutional layers to 3×3, which helps to eliminate noise in the feature and improve the abstract representation of the feature, followed by a pooling layer. The fourth convolutional layer, called the feature advanced presentation layer, helps eliminate redundant features and improve representative features, followed by a pooling layer. Three-layer fully connected layer for feature classification and parameter optimization during backpropagation. At the end of the network, the Softmax layer is used for classification. Softmax is a supervised learning method for multi-classification problems [33, 34] that provides important confidence levels for classification, where 0 indicates the lowest confidence and 1 indicates the highest.

3.1. The k-means++ clustering algorithm extracts the dominant color features

The k-means++ clustering algorithm is a 'hard clustering' algorithm, in which each sample in a dataset is scored 100% into a category. In contrast, 'soft clustering' can be understood as a certain probability that each sample is sorted into a certain category.

The original k-means algorithm initially randomly selects k points in the dataset as cluster centers, while k-means++ selects k cluster centers according to the following idea: assuming that n initial cluster centers have been selected.

The k-means++ clustering algorithm [35] is described as follows:

1. Randomly select a sample from the dataset as the initial cluster center c_i.
2. First calculate the shortest distance between each sample and the current existing cluster center (that is, the distance from the nearest cluster center), expressed in terms of representation $D(x)$; The probability that each sample will be chosen as the next cluster center is then calculated $\frac{(D(x))^2}{\sum_{x \in X}(D(x))^2}$. Finally, the next cluster center is selected according to the roulette method.
3. Repeat Step 2 until a common cluster center k is selected.
4. For each sample $x_(i)$ in the dataset, calculate its distance to a cluster center k and divide it into classes corresponding to the cluster center with the smallest distance.
5. For each category $c_(i)$, recalculate its cluster center $c_i = \frac{1}{|c_i|} \sum_{x \in c_i} X$ (i.e. the centroid of all samples belonging to that class).
6. Repeat Steps 4 and Step 5 until the position of the cluster center no longer changes.

3.2. Deep multi-feature fusion

Convolutional layers, nonlinearly activated transformations, and pooled layers are the three basic components of CNNs. By superimposing multiple convolutional layers with nonlinear operations and multiple pooling layers, a deep CNNs can be formed, which extract input features in layers, with invariance and robustness [36]. With specific architectures, such as local connections and shared weights, CNNs tend to have good generalization capabilities. Convolutional layers with nonlinear operations [32] are as follows (1):

$$x_j^l = f\left(\sum_{i=1}^{M} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

Where the matrix x_i^{l-1} is the ith feature map of the $l-1$ layer, x_j^l is the jth feature map of the current layer l, and M is the number of input feature maps. k_l^{ij} and b_j^l are randomly initialized and set to zero, then fine-tuned precisely by backpropagation. $f(\cdot)$ is a nonlinear activation function, and $*$ is a convolution operation.

The denoising depth convolution features f_c of the network structure output and the main color features extracted by the k-means++ clustering algorithm f_l are constructed according to the cascading and features of (2) to construct the deep integration features f_m.

$$f_m = \alpha f_c + \beta f_l \tag{2}$$

Due to the high characteristic dimensions and limited training samples, overfitting is a serious problem that can occur. To solve this problem, the Dropout [37] method was used, which is a method of randomly deleting neurons during learning in Figure 2. During training, every time the data is passed, the neurons in the hidden layer are randomly selected, and then deleted, and the deleted neurons no longer transmit signals; During the test, although all neuronal signals are transmitted, the output of each neuron is multiplied by the deletion ratio at the time of training before the output. Therefore, the deleted neurons are not involved in forward transmission and are no longer used in the backward propagation process. At different training stages, deep networks form different neural networks by randomly discarding neurons. The Dropout method prevents complex co-adaptations, and neurons can learn more correct features.



Figure 2: MCNN+ configuration

4. Experiment

4.1. Datasets

To verify the model orientation in this paper, a dataset of gemstone images obtained from [38] was selected for analysis. The dataset contains more than 3,200 images of different gemstone. The images are divided into 87 categories, which have been divided into training data and test data. Some of these examples are shown in Figure 3. Images are obtained under very different lighting and background color conditions. A total of 2800 images were used for training and 400 images were reserved for testing. For each class, there are 24-44 training images and 4-5 test images available.

The Epoch of the CNNs network model with deep multi-feature fusion is set to 50, the number of iterations is 8, and the batch size is 1000 images. The convolutional kernel size of the first convolutional layer is 5 *5, and the filter is 32; The convolutional kernel size of the second convolutional layer is 5*5, and the filter is 64; The convolutional kernel size of the third and fourth convolutional layers is 3*3, and the filter is 128. The convolution step is set to 1 and the fill is set to 0. The pooling layer step size is set to 2, and the pooling window size is 2*2. Finally, the Softmax layer has 10 neural units, indicating that the images are divided into 10 classes.



Figure 3: Partial image of the gemstone image dataset

The optimal solution is obtained by minimizing the loss function. This study uses the cross-entropy error function [14] as the loss function, which is expressed as (3):

$$Q = -\frac{1}{N}\sum_{n=1}^{N}\left((y)_n (log)_2 ((o)_n) + (\tilde{y})_n (log)_2 \left(\tilde{o_n}\right)\right) \quad (3)$$

Where the $\tilde{y_n} = 1 - y_n, \tilde{o_n} = 1 - o_n$, Nnumber of samples is trained for batch processing, y is the true label value for each sample, and o is the actual output value of the network.

To optimize the loss function, an optimization method is required, and the Adam optimizer is used for the experiment [15]. Combine the advantages of two optimization algorithms, AdaGrad and RMSProp. Considering the first-order moment estimation (the mean of the gradient) and the second-order moment estimation (the undercentric variance of the gradient) of the gradient, the update step is calculated.

4.2. *Experimental results and analysis*

4.2.1. *Evaluation function*

To evaluate the performance of the algorithm, this paper uses a training validation test scheme in which 80% of the data is used as the training set and 20% as the test set. The test accuracy rate is used to evaluate the performance of the model and is expressed as (4):

$$Accuracy = \frac{R}{T} \quad (4)$$

where: R is the sample with the correct classification and T is the total sample.

4.2.2. *Experimental results*

In this study, a convolutional neural network that deeply integrates the main color features with position and shape-related features, referred to as MCNN+. The model training results of MCNN+ are shown in the following Figure 4 and Figure 5. As can be seen from the Figure 4, the model gradually reaches a steady state after iterative training. One of the best classifications has an accuracy rate of 82%.

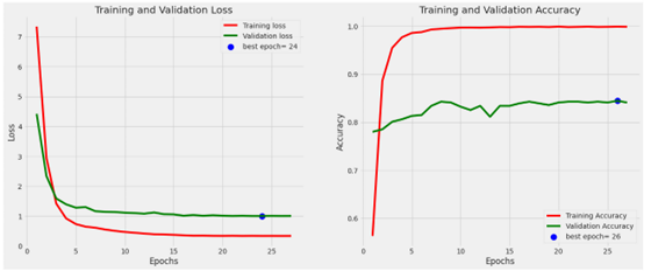


Figure 4: Model training results

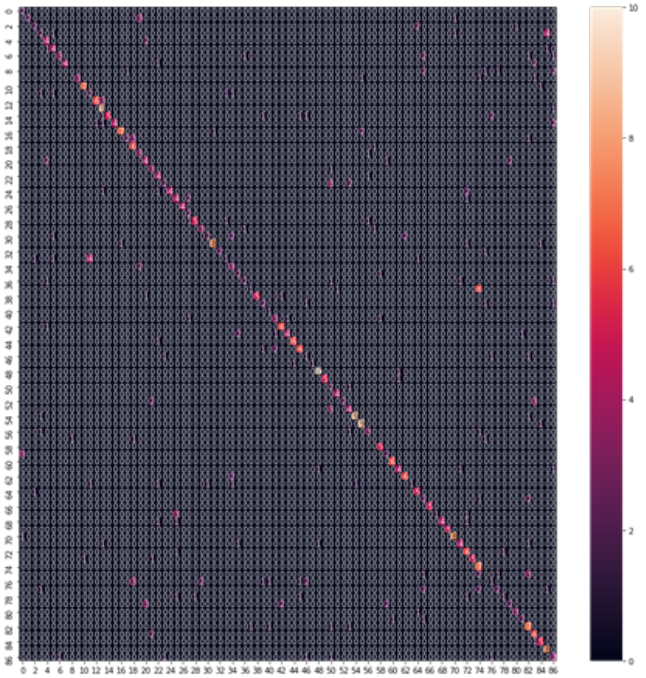


Figure 5: Confusion matrix for model classification

4.2.3. *Model Comparison*

We also compares the MCNN+ model with several machine learning models and deep learning models, as shown in the following Table 1.

Table 1: Comparison of different model results

Methods	Accuracy
CNN	0.721
Random Forest	0.694
Logistic Regression	0.687
SVM	0.669
Naive Bayes	0.627
InGG16	0.732
ResNet50	0.746
MCNN+	0.810

As can be seen from the table, the model in this paper can achieve the best classification results when classifying the gemstone image data set. And it can be improved by nearly 9% compared to the CNN model.

5. Conclusion

In this study, we proposed a CNN gemstone image classification algorithm with deep multi-feature fusion, which effectively integrates the main color characteristics of gemstone images with the spatial position and shape characteristics extracted by convolutional neural networks. This approach addresses the limitation of traditional CNN models, which primarily focus on spatial position information and are less sensitive to color information. Experimental results demonstrate that MCNN+ not only significantly improves image classification performance but also exhibits greater stability and robustness. Additionally, our experiments show that different weight values in the MCNN+ network models have varying effects on experimental performance.

In future research, we aim to implement adaptive weight mechanisms to further enhance the model's performance. Additionally, we plan to explore the integration of other feature extraction techniques, such as texture and edge detection, to enrich the feature set used for classification. We will also investigate the application of this model to other types of images and domains to test its versatility and generalizability. Finally, leveraging advanced techniques like transfer learning and self-supervised learning could further optimize the model and reduce the reliance on large labeled datasets.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment College of Jewelry, Shanghai Jian Qiao University.

References

[1] D. Bersani, P. P. Lottici, "Applications of raman spectroscopy to gemology", *Analytical and bioanalytical chemistry*, vol. 397, pp. 2631–2646, 2010, doi:10.1007/s00216-010-3700-1.

[2] J. E. Shigley, "A review of current challenges for the identification of gemstones", *Geologija*, , no. 64, 2008, doi:10.2478/v10056-008-0048-8.

[3] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions", *Journal of big Data*, vol. 8, pp. 1–74, 2021, doi:10.1186/s40537-021-00444-8.

[4] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, "Review on convolutional neural networks (cnn) in vegetation remote sensing", *ISPRS journal of photogrammetry and remote sensing*, vol. 173, pp. 24–49, 2021, doi:10.1016/j.isprsjprs.2020.12.010.

[5] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, M. Unser, "Cnn-based projected gradient descent for consistent ct image reconstruction", *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1440–1453, 2018, doi:10.1109/TMI.2018.2832656.

[6] H. Lee, H. Kwon, "Going deeper with contextual cnn for hyperspectral image classification", *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, 2017, doi:10.1109/TIP.2017.2725580.

[7] I. Maula, V. Amrizal, A. H. Setianingrum, N. Hakiem, "Development of a gemstone type identification system based on hsv space colour using an artificial neural network back propagation algorithm", "International Conference on Science and Technology (ICOSAT 2017)-Promoting Sustainable Agriculture, Food Security, Energy, and Environment Through Science and Technology for Development", pp. 104–109, Atlantis Press, 2017, doi:10.2991/icosat-17.2018.24.

[8] A. Ostreika, M. Pivoras, A. Misevičius, T. Skersys, L. Paulauskas, "Classification of objects by shape applied to amber gemstone classification", *Applied Sciences*, vol. 11, no. 3, p. 1024, 2021, doi:10.3390/app11031024.

[9] A. Ostreika, M. Pivoras, A. Misevičius, T. Skersys, L. Paulauskas, "Classification of amber gemstone objects by shape", 2020, doi:10.20944/preprints202008.0336.v1.

[10] R. S. Castro Rios, "Researching of the deep neural network for amber gemstone classification", Master's thesis, Universitat Politècnica de Catalunya, 2018.

[11] S. Sinkevičius, A. Lipnickas, K. Rimkus, "Multiclass amber gemstones classification with various segmentation and committee strategies", "2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)", vol. 1, pp. 304–308, IEEE, 2013, doi:10.1109/IDAACS.2013.6662694.

[12] X. Liu, J. Mao, "Research on key technology of diamond particle detection based on machine vision", "MATEC Web of Conferences", vol. 232, p. 02059, EDP Sciences, 2018, doi:10.1051/matecconf/201823202059.

[13] B. H. Y. Chow, C. C. Reyes-Aldasoro, "Automatic gemstone classification using computer vision", *Minerals*, vol. 12, no. 1, p. 60, 2021, doi:10.3390/min12010060.

[14] K. Hurrell, M. L. Johnson, *Gemstones: a complete color reference for precious and semiprecious stones of the world*, Chartwell Books, 2016.

[15] C. M. Breeding, A. H. Shen, S. Eaton-Magaña, G. R. Rossman, J. E. Shigley, A. Gilbertson, "Developments in gemstone analysis techniques and instrumentation during the 2000s.", *Gems & Gemology*, vol. 46, no. 3, 2010.

[16] R. Liddicoat, "Developing the powers of observation in gem testing", *Gems Gemol*, vol. 10, pp. 291–319, 1962.

[17] E. Fritsch, C. M. Stockton, "Infrared spectroscopy in gem identification", *Gems & gemology*, vol. 23, no. 1, pp. 18–26, 1987.

[18] A. L. Jenkins, R. A. Larsen, "Gemstone identification using raman spectroscopy", *disclosure*, vol. 7, p. 9, 2004.

[19] L. Kiefert, S. Karampelas, "Use of the raman spectrometer in gemmological laboratories", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 80, no. 1, pp. 119–124, 2011, doi:10.1016/j.saa.2011.03.004.

[20] T. He, "The applications of ultraviolet visible absorption spectrum detection technology in gemstone identification", "Proceedings of the 5th International Conference on Materials Engineering for Advanced Technologies (ICMEAT 2016), Quebec, QC, Canada", pp. 5–6, 2016.

[21] S. Thompson, F. Fueten, D. Bockus, "Mineral identification using artificial neural networks and the rotating polarizer stage", *Computers & Geosciences*, vol. 27, no. 9, pp. 1081–1089, 2001, doi:10.1016/S0098-3004(00)00153-9.

[22] N. A. Baykan, N. Yılmaz, "Mineral identification using color spaces and artificial neural networks", *Computers & Geosciences*, vol. 36, no. 1, pp. 91–97, 2010, doi:10.1016/j.cageo.2009.04.009.

[23] H. Izadi, J. Sadri, M. Bayati, "An intelligent system for mineral identification in thin sections based on a cascade approach", *Computers & Geosciences*, vol. 99, pp. 37–49, 2017, doi:10.1016/j.cageo.2016.10.010.

[24] H. Pereira Borges, M. S. de Aguiar, "Mineral classification using machine learning and images of microscopic rock thin section", "Advances in Soft Computing: 18th Mexican International Conference on Artificial Intelligence, MICAI 2019, Xalapa, Mexico, October 27–November 2, 2019, Proceedings 18", pp. 63–76, Springer, 2019, doi:10.1007/978-3-030-33749-0_6.

[25] J. Maitre, K. Bouchard, L. P. Bédard, "Mineral grains recognition using computer vision and machine learning", *Computers & Geosciences*, vol. 130, pp. 84–93, 2019, doi:10.1016/j.cageo.2019.05.009.

[26] Y. Zhang, M. Li, S. Han, Q. Ren, J. Shi, "Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms", *Sensors*, vol. 19, no. 18, p. 3914, 2019, doi:10.3390/s19183914.

[27] K. Muthukaruppan, S. Thirugnanam, R. Nagarajan, M. Rizon, S. Yaacob, M. Muthukumaran, T. Ramachandran, "A comparison of south east asian face emotion classification based on optimized ellipse data using clustering technique", *Journal of Image and Graphics*, vol. 3, no. 1, 2015, doi:https://doi.org/10.18178/joig.3.1.1-5.

[28] Y. Karayaneva, D. Hintea, "Object recognition in python and mnist dataset modification and recognition with five machine learning classifiers", *Journal of Image and Graphics*, vol. 6, no. 1, pp. 10–20, 2018.

[29] L. Wang, B. Liu, S. Xu, J. Pan, Q. Zhou, "Ai auxiliary labeling and classification of breast ultrasound images", *Journal of Image and Graphics*, vol. 9, no. 2, pp. 45–49, 2021.

[30] N. M. Trieu, N. T. Thinh, "A study of combining knn and ann for classifying dragon fruits automatically", *Journal of Image and Graphics*, vol. 10, no. 1, pp. 28–35, 2022.

[31] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, "Recent advances in convolutional neural networks", *Pattern recognition*, vol. 77, pp. 354–377, 2018, doi:10.1016/j.patcog.2017.10.013.

[32] Y. Chen, C. Li, P. Ghamisi, X. Jia, Y. Gu, "Deep fusion of remote sensing data for accurate classification", *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1253–1257, 2017, doi:10.1109/LGRS.2017.2704625.

[33] R. Kiran, P. Kumar, B. Bhasker, "Dnnrec: A novel deep learning based hybrid recommender system", *Expert Systems with Applications*, vol. 144, p. 113054, 2020, doi:10.1016/j.eswa.2019.113054.

[34] X. Peng, X. Zhang, Y. Li, B. Liu, "Research on image feature extraction and retrieval algorithms based on convolutional neural network", *Journal of Visual Communication and Image Representation*, vol. 69, p. 102705, 2020, doi:10.1016/j.jvcir.2019.102705.

[35] D. Arthur, S. Vassilvitskii, "k-means++: The advantages of careful seeding", Tech. rep., Stanford, 2006.

[36] Y. Bengio, A. Courville, P. Vincent, "Representation learning: A review and new perspectives", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013, doi:10.1109/TPAMI.2013.50.

[37] D. Steinkraus, I. Buck, P. Y. Simard, "Using gpus for machine learning algorithms", "Eighth International Conference on Document Analysis and Recognition (ICDAR'05)", pp. 1115–1120, IEEE, 2005, doi:10.1109/ICDAR.2005.251.

[38] "Gemstone images", https://www.kaggle.com/lsind18/gemstone-images.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



HAOYUAN HUANG has done his bachelor's degree in Computer Art Design from Zhongnan University of Economics and Law in 2008. He has done her master's degree in Digital Media and Game Design from Wuhan University in 2010.

He is currently a lecturer at the School of Jewelry, Shanghai Jianqiao University, focusing her research on jewelry design in the field of product design.



RONGCHENGCUI has done his bachelor's degree in Computer Science and Technology from Shijiazhuang Tiedao University in 2019. He has done his master's degree in Electrical Information Technology from Shanghai University of Electric Power in 2022.

He is currently an engineer at Microsoft China. His main research areas include machine learning, image processing, and computer vision.

Received: 07 August, 2024, Revised: 05 September, 2024, Accepted: 06 September, 2024, Online: 15 September, 2024

DOI: https://doi.org/10.55708/js0309002

The Dual Impact of AI in Clinical Trials: Perspective

Muhammad Yaqub, Lan He

School of Biomedical Sciences, Hunan University, Changsha, 410000, P.R China

*Corresponding author: Muhammad Yaqub, Hunan University, myaqub@hnu.edu.cn

ABSTRACT: The use of artificial intelligence in clinical trials opens the way for a period of transformation in which the potential to enhance efficiency, precision, and scope in clinical investigation is enormous. In this perspective, the promise and perils of AI regarding clinical trials are critically reviewed. On one hand, it's where AI can really make some revolutionary changes to the most critical aspects of trial design and execution—like patient recruitment, data management, and predictive analytics—which are now adaptive and more driven by data; on the other hand, AI applied in clinical trials also brings considerable problems related to data integrity, ethical considerations, and regulatory compliance. Such potential of AI to further biases, break patient confidences, and exacerbate already wide inequities in healthcare raises the need for vigilant oversight. As AI is rapidly evolving, so too are principles that guide transparency, fairness, and ethical rigors that guide their application in clinical trials. This work presents a case for strong AI frameworks that are subject to firm validation and ethical scrutiny, which will ensure that the benefits of AI are realized with reduced risks associated with it. If the field goes down this road, AI integration in clinical trials is going to be the real catalyst for innovation, which could translate into achieving better patient outcomes and expanding frontiers for medical science.

KEYWORDS: Artificial Intelligence, Clinical Trials, Patient Outcomes, Predictive Analytics, Healthcare Inequities

1. Introduction

One of the most extreme developments in medical research is the use of AI in clinical trials. Already available are emerging AI technologies offering enormous potential for efficiency, accuracy, and scalability in clinical trials.[1–3] . That promise is wrapped in a series of major challenges and key ethical concerns. There are many risks associated with implementing AI within clinical trials. Serious issues of data privacy, bias in algorithms, transparency, and the possible increase of inequities in health characterize a very difficult area of navigation. It also brings along several questions on the level of regulatory compliance, validity of the AI-driven outcomes, and whether healthcare is prepared to absorb full potential for AI while protecting the rights and well-being of the patient [4,5]. As summarized in Table 1, the literature highlights the transformative impact of AI on clinical trials, particularly in improving patient recruitment efficiency, enhancing data management, and addressing ethical and privacy concerns. These studies underscore both the potential benefits and the challenges that need to be navigated to fully integrate AI into clinical research.

It is therefore advanced in this paper that the dual-edged potential of AI in the context of clinical trials demands an appropriate and integrative perspective of the optimistic possibilities and potential risks concomitant with it. Current advancements in this area, related ethical considerations, and regulatory challenges were looked at in order to provide a balanced view in respect of the optimization of clinical trials using AI and the critical issues that have to be tackled for its successful integration. It contributes to the discussion on responsible and effective application of AI about medical research, and subsequently toward an innovation-sound and ethical future for AI-aided clinical trials. Figure 1 illuminating the promising role of AI in healthcare.

Table 1: Key Literature on the Integration of Artificial Intelligence in Clinical Trials

Reference	Year	Title	Key Findings	Relevance to AI in Clinical Trials
[6]	2019	High-performance	AI has the potential to significantly	Discusses the overall potential of

		medicine: the convergence of human and artificial intelligence	enhance the accuracy, efficiency, and personalization of healthcare, including in clinical trials.	AI in healthcare, highlighting its transformative impact on clinical trials.
[7]	2020	The future of digital health with federated learning	Federated learning offers a way to leverage AI while preserving patient data privacy, a critical aspect of clinical trials.	Provides insights into maintaining data privacy in AI-driven clinical trials, a key ethical concern.
[8]	2019	Optimizing patient recruitment with AI	AI can streamline patient recruitment by analyzing electronic health records to identify candidates more efficiently.	Focuses on how AI can address the significant challenge of patient recruitment in clinical trials.
[9]	2020	International evaluation of an AI system for breast cancer screening	Demonstrates the accuracy of AI in medical imaging, with potential applications in clinical trial monitoring and outcome prediction	Illustrates the practical applications of AI in diagnostic processes relevant to clinical trials.
[10]	2021	Overview of deep learning applications in medical image analysis	Reviews the state-of-the-art deep learning methods in medical image analysis, highlighting their potential in disease diagnosis and clinical trials.	Provides a comprehensive review of AI applications in medical imaging, relevant for trials involving imaging-based outcomes.
[11]	2020	Ethical considerations in the deployment of AI in healthcare	Discusses the ethical implications of AI in healthcare, including issues of bias, transparency, and accountability.	Essential for understanding the ethical challenges AI presents in the clinical trial context.
[12]	2017	Dermatologist-level classification of skin cancer with deep neural networks	AI achieved dermatology-level accuracy in skin cancer diagnosis, suggesting AI's potential to improve diagnostic accuracy in trials.	Demonstrates the effectiveness of AI in diagnostics, relevant for trials focusing on dermatological conditions.
[13]	2017	Artificial intelligence in healthcare: past, present, and future	Explores the development of AI in healthcare and its potential future applications, including clinical trials.	Provides historical context and future outlook on AI's role in clinical trials.
[14]	2019	The role of artificial intelligence in precision medicine	AI enables precision medicine approaches in clinical trials, allowing for more tailored treatment interventions.	Highlights how AI can enhance the personalization of clinical trials through precision medicine.
[15]	2018	How artificial intelligence is changing	AI accelerates the drug discovery	Discusses AI's role in the early stages of clinical

		drug discovery	process, including in	trials, particularly in drug discovery and testing.

2. The Impact of AI in Medical Imaging

The effect of AI on clinical trials, even more so in the analysis of medical imaging, goes deep, as it helps in improving image analysis, smoothing trial workflows, and enhancing diagnostic precision. Speaking precisely, some of the ways AI can modify medical imaging in clinical trials include:

2.1. Automated Image Analysis and Interpretation

Medical imaging modalities, including but not limited to MRI, CT, and PET, are increasingly used in clinical trials for assessment of diseases process, treatment efficacy assessment, or biomarkers evaluation [16,17]. Traditional radiologist interpretation of images is very time-consuming and may be variable. Algorithms based on AI, especially those using deep learning techniques, can be programmed to look at medical imagery for patterns indicating lesions or other abnormalities that may well be too slight for the human eye to recognize. For example:

- AI-powered tools in an oncology trial can assist by automatic detection and quantification of tumor size, shape, and growth patterns in MRI or CT images, improving the assessment of treatment responses.
- AI in cardiology can analyze echocardiograms or cardiac MRIs to measure heart function, such as ejection fraction, wall motion, and other variables that provide insight into drug effects on cardiovascular health.

AI-driven automation decreases interpretation time for images, enhances consistency in the review of images, and decreases the probability of human-related errors. The outcome is speedier trial results and more valid results, as the imaging biomarkers could be studied more precisely.

2.2. Enhanced Image-Based Biomarker Discovery

Biomarkers are important in a clinical trial in monitoring disease process or treatment response [18]. For imaging, this could involve the detection of slight structural or functional alterations in tissues, which are indicative of the disease. Artificial Intelligence has this phenomenal capability to discover new imaging biomarkers by analyzing huge datasets, with subtle patterns that may evade human eyes [19]. For instance:

- So far, these models have been applied in neurology for brain MRI reviewing, considering a number of biomarkers for various diseases, such as Alzheimer's disease and multiple sclerosis [20], which are grounded on different brain volume changes, cortical thickness, and white matter integrity.
- AI will find the radiomic features in oncology, such as microscopic variations either in tumor textures or shapes related to variations in patient outcomes, allowing early treatment efficacy predictions.

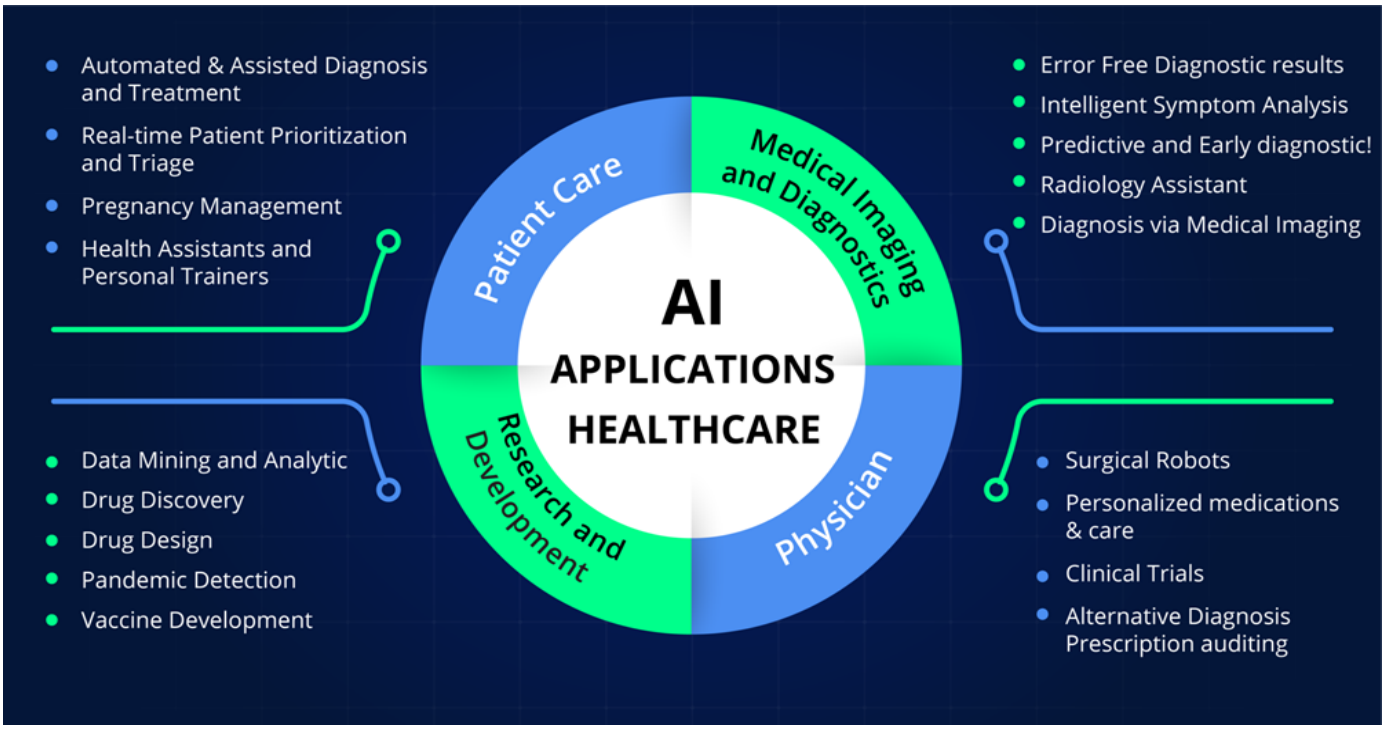


Figure 1: The promising role of AI in Healthcare

AI, therefore, applies to an improved ability in detection and quantification of complex biomarkers from imaging information, which enables better prediction of the progression of diseases and responses to pharmacological treatments. Consequently, this process has encouraged the formulation of more effective therapies and personalized ones during clinical trials, out of which comes the shortening of timelines for bringing efficacious therapies to market.

2.3. Real-Time Image Monitoring in Adaptive Trials

AI represents the backbone of adaptive clinical trials, most seen in real-time image monitoring during trials, for which imaging is an enabler of response assessment in patients and modifications to trials [21]. Some decisions regarding dosage, continuing or modifying treatment, or even cohort adjustments depend mostly on scheduled reviews. AI can give data-driven insights at the exact moment by analyzing real-time imaging results. For example, in oncology trials, the AI system can track day-to-day changes in tumor size, density, and vascularity from MRI or CT scans continuously, which, based on a poor response of a patient, clinicians can alter the regimen of treatment immediately. This will save precious time in making critical adjustments, therefore offering the best possible care for a patient. AI-driven image analysis also cuts down on human error; hence, allowing more consistent and objective assessments across large cohorts of patients. AI makes clinical trials more flexible, meaning that faster and more precise assessments will enable the creation of better treatments.

2.4. AI in Image-Based Inclusion/Exclusion Criteria

It often happens that medical imaging itself is crucial for the inclusion or exclusion of patients in clinical trials, depending on whether the extent of the disease, organic damage, or size of lesions meets the particular inclusion or exclusion criteria. Conventionally, radiologists would review such criteria manually by going through the imaging data; this is a very time-consuming process and may be prone to inaccuracies and inconsistencies. Artificial Intelligence solves this through automation of involved processes. It ensures the screening of the patients systematically and precisely based on the eligibility criteria brought about by the trial in mention [22].

For example, in neurological clinical trials, algorithms in AI can independently analyze MRI images to look for symptoms of brain atrophy or even lesions that will determine if a patient can be included or not. Similarly, AI in oncology research is able to evaluate the dimensions and dissemination of tumors, classifying patients into the stages of investigation accordingly. This automation simplifies the process of screening, enhances the accuracy in patient selection, and thereby accelerates participants' recruitment, which has been a major bottleneck in many clinical trials.

2.5. Improved Patient Monitoring through Longitudinal Imaging

Longitudinal imaging [23] is very often necessary in many clinical trials, especially those that target chronic or progressive diseases; this usually requires the repeated scanning of patients over time to monitor changes in disease progression or therapeutic effects. AI enhances this process by analyzing sequential imaging data and pointing out subtle changes that may pass undetected by the human eye. For example, in neurodegenerative disease trials, AI can quantify the change in brain volume between time points to give early signs of disease progression or response to therapy, while in oncology trials, the rate of tumor shrinkage or detection of new metastases in follow-up scans from AI can continuously provide feedback on treatment efficacy. The potential to automatically analyze complex imaging data over long periods ensures increased precision in monitoring, allowing for adjustments in trials based on real patient responses and therefore enabling rapid and true assessments of the efficacy of drugs.

2.6. Reducing Variability in Imaging Interpretations

One major challenge facing clinical trials involving imaging modalities is the variation in the interpretation of the same image by different radiologists. Such variability leads to heterogenous data that might affect the outcome of such research. Artificial intelligence applications overcome this problem by applying uniform criteria to all image analyses, hence enabling uniform interpretation of images irrespective of the trial site or the personnel involved.

While artificial intelligence systems can, for example, measure such characteristics of lesions objectively for conditions like multiple sclerosis or cancers where the quantity and dimensions of lesions are a primary critical endpoint in clinical trials on all subjects. This permits a reduction of inter-reader variability and thus provides an assessment platform that is both standardized and reliable. The consistency of image reading is particularly essential in multisite clinical trials, where the results of image reading coming from different sites should be comparable. By mitigating the variability of humans,

artificial intelligence enhances the robustness and reliability of trial data, leading to a more exact conclusion.

2.7. AI for Enhanced Image-Guided Drug Delivery

The use of imaging in delivering drugs to a specific part of the body during most clinical targeted therapies, especially in cancer treatment, is critical. Artificial intelligence enhances the precision in such analyses of image data and hence could allow therapeutic agents to reach the sites with greater accuracy. The examples are that AI, in radiotherapy research, can analyze CT or MRI scans to delineate tumor margins with high precision so that radiation is delivered onto cancerous tissues while sparing other healthy ones. For example, in trials with nanoparticle or antibody-drug conjugates, AI can aid in determining optimum delivery sites and/or real-time drug dispersion with companion imaging techniques like PET or SPECT. Greater precision reduces the risk of unwanted off-target effects, improves the overall safety profile of these therapeutic interventions, and enhances the possibility of successful realization of the intended outcomes of the therapeutic interventions themselves. While AI-driven imaging guidance in drug delivery serves the dual purpose of enhancing effectiveness, the most important part is that it optimizes trial protocols by enhancing efficiency and productivity within clinical research [24].

3. Case Studies

The following are some case studies that give an idea of the realistic implementation of AI in clinical trials. It underlines and identifies real practical examples in order to explain exactly how AI impacted different stages of clinical research.

3.1. IBM Watson for Drug Discovery – Accelerating Cancer Research

IBM Watson for Drug Discovery has been employed by various pharmaceutical companies and research institutions to accelerate the identification of potential therapeutic targets and biomarkers for diseases like cancer. Working with the Barrow Neurological Institute, Watson [25] uncovered five new ALS-related gene candidates in less than two months' work that, if done by human researchers, would have taken years. Scanning through thousands of scientific papers and clinical trial records, AI was able to quickly create links between genes and ALS, accelerating the pathway to finding new medications.

The AI-driven literature and data mining of Watson reduced the time it took to sift through datasets that were too complex; this hastened the early phases of clinical trials and hypothesis generation. This case underlines the capability of AI in manipulating large volumes of unstructured data, hence reinforcing trial design and biomarker identification, especially in complex diseases like ALS and cancer.

3.2. Tempus – Personalized Oncology Trials

Tempus is a Chicago-based technology company that uses AI to study clinical and molecular data to personalize cancer treatment. Its platform collates clinical data, including the genetic profile of patients, in order to match them with suitable clinical trials.

Tempus [26] has partnered with the University of Chicago Medicine to incorporate its AI platform into clinical trial matching at the hospital. In return, it analyzes the genomic sequencing data and clinical records to identify appropriate clinical trials for cancer patients and help match the right patients into the relevant oncology trials. This reduced the time taken by the system in matching patients with trials, while the trials the patients were being subjected to were more accurate. AI-driven patient matching works to improve enrollment rates, normally bottlenecks to oncology research. Additionally, this will enable the improvement of prospects for therapeutic success on account of the closer genetic profile alignment of a patient with trial parameters for more tailored and effective treatments.

3.3. Medidata – AI for Virtual Clinical Trial

Medidata is a company focused on cloud-based solutions for life sciences, having come up with an AI platform that allows the implementation of virtual clinical trials. The approach thereby minimizes physical visits and allows for remote data collection, which applies to rare diseases or pandemics, such as COVID-19. Medidata [27] conducted virtual trials of a candidate COVID-19 vaccine by implementing AI for remote monitoring of patient data. The AI-enabled platform collected data from wearables and patient-reported outcomes, while making the processing easier with a reduced administrative burden. This was interpreted as monitoring patients in real-time, without asking patients to travel to trial sites.

Virtual trials by Medidata removed geographical boundaries and allowed more patients to participate in parts of the world where access to clinical trial sites was at a premium. The integration of data from wearables enabled by AI also improved monitoring of patient health metrics, thus allowing quicker safety assessments and flexibility in the management of trials. That has been

particularly useful during the pandemic for running trials in a decentralized manner and finds wider application now.

3.4. AI in Adaptive Clinical Trials – Novartis and Microsoft

Novartis has allied with Microsoft to use AI in the design of adaptive clinical trials where real-time changes can be made to trial protocols based on intermediate outcomes [28]. This becomes particularly important in oncology, where patient responses can be highly variable. AI models, developed under the partnership between Novartis and Microsoft, predicted optimal dosing levels and signals of early treatment efficacy. For instance, AI models spotted subgroups of more responsive patients during a recent breast cancer trial, allowing for adaptive changes in the protocol of this trial focused on these subgroups.

Adaptation in trials-in fact, making changes according to responses-allows for efficiency in research and safety for the patient. AI enhances that ability to dynamically alter trial parameters, reducing time and cost it takes in new treatments. For example, in the case of this AI-driven trial on breast cancer, the number of patients exposed to ineffective treatments was reduced while hastening the approval process of successful therapies.

3.5. Benevolent AI – Accelerating Drug Repurposing

Benevolent AI applies the power of AI to find new drug candidates and indications that find new uses for already-approved drugs [29]. It reads through biomedical data like clinical trials, patents, and scientific literature to identify new potential uses for already-approved drugs. In the COVID-19 pandemic, Benevolent AI mentioned the use of baricitinib-a rheumatoid arthritis medication-as a likely candidate against COVID-19. The AI system curbed through mountains of data and suggested that the drug could reduce the cytokine storm that happens in serious cases of COVID-19, and hence it is an ideal candidate for repurposing.

Clinical trials started for baricitinib, and it proved very beneficial for COVID-19 patients. The drug was able to garner use authorization in emergency situations from the FDA. This is just one example in which AI accelerates a process that, without AI, involves tremendous time and resources: finding new uses for existing drugs. AI analyzes cross-domain datasets, driving much higher efficiency in drug repurposing.

3.6. Verge Genomics – AI for ALS Clinical Trials

Verge Genomics employs AI and machine learning to map the human genome for new treatments of neurodegenerative diseases such as ALS [30]. In collaboration with the Cedars-Sinai Medical Center, Verge has fielded its AI platform against ALS patient data and surfaced novel targets for potentially informed clinical trials. In this respect, by applying AI to identify genetic patterns, Verge can accelerate the preclinical stages and move promising candidates into clinical trials more rapidly than would have otherwise been possible. The AI platform at Verge automated target identification and shaved huge margins off ALS drug development times and costs. The rapid pace from identification to clinical trials punts on AI's power for accelerating the initial parts of neurodegenerative disease clinical research, where advances usually take so much time.

These case studies epitomize how AI can transform everything from patient recruitment and data analysis to adaptive trial designs and drug repurposing within the clinical trial space. AI-driven tools enhance clinical research speed, efficiency, and precision-enabling better outcomes in oncology, neurodegenerative diseases, and pandemic response efforts. These examples demonstrate great strides that have been taken in real-world uses where AI makes the clinical trial more efficient, personalized, and accessible.

4. The Promise of AI in Clinical Trials

Clinical trials powered by AI can bring about a wealth of advantages to almost flip a new leaf in medical research and treatment. This starts from the patient selection stage and takes the shapes of the analysis, monitoring, and outcome prediction for clinical trials. document.

Table 2: Summary of the Promises and Perils of AI in Clinical Trials

Aspect	Promises	Perils
Efficiency [31]	Streamlines data analysis and patient selection.	- Over-reliance on AI could lead to overlooking nuanced clinical insights.
Accuracy [32]	- AI models can process large large datasets, identifying patterns beyond human capability.	- Potential for bias in AI algorithms due to biased training data.
Cost-Effectiveness [33]	- Reduces costs by automating routine tasks and improving trial design.	- High initial costs for AI implementation and integration.

Patient Stratification [25]	- Identifies optimal patient subgroups for targeted therapies.	- Risk of excluding underrepresented groups, leading to inequitable outcomes.
Speed [34]	- Accelerates the timeline of trials by predicting outcomes and optimizing processes.	- Speed at the expense of thoroughness may compromise the trial's integrity.
Regulatory Compliance [35]	- Assists in meeting regulatory requirements by maintaining accurate records and analysis.	- Regulatory bodies may struggle to keep pace with rapidly evolving AI technologies.
Ethical Considerations [36]	- AI can help ensure ethical trial designs by identifying potential issues in advance.	- Ethical dilemmas around data privacy, consent, and the use of AI in decision-making.
Interpretability [37]	- Enhances understanding of complex data through advanced analytics and visualization tools.	- Black-box nature of some AI models may hinder transparency and trust in AI-driven decisions.

4.1. Enhanced Patient Recruitment and Retention

Patient recruitment and retention are the biggest challenges to clinical trials, namely, enrolling eligible patients or participants. The traditional model of doing things relies on manual efforts, which become time-consuming, prone to mistakes, and cumbersome. AI can fast-track patient recruitment by analyzing voluminous electronic health records and other medical data for suitable candidates with greater speed and accuracy. AI Algorithms can thus be trained to provide an estimate of the patient retention probability by detecting those factors that could influence a participant's likelihood of completing a trial, thus improving retention rates and overall success of a trial.

4.2. Improve how one manages and analyzes data

Clinical trials involve huge volumes of data, from patient demographics to clinical measurements, genomic data, and image data. Management and analysis operations for this kind of data are bulky and resource intensive. AI has the potential to innovate data management through automating means of processing and analyzing big datasets to researchers for them to find out hidden patterns and insights that, otherwise, would not be evident in the use of conventional statistical methods. Through machine learning algorithms, one is able to come up with the correlations, predict the outcome, and even hypothesize new hypotheses that enable researchers to fast track the processes in research activities, thereby increasing the accuracy of results realized in trials.

4.3. Personalized Medicine and Adaptive Trials

AI also deepens analysis of datasets in ways that foster the development of personalized medicine approaches within clinical trials. In this regard, AI would be useful in developing treatment trials through integrating genomic, phenotypic, and clinical data in tailoring treatments to the characteristics of the individual patient. This customization will maximize the possibilities of success of the treatment while minimizing its side effects. Moreover, AI can support adaptive designs whereby the parameters of the trial are adapted based on the results of the interim analysis. A very key implication of adaptability is running quite a lot more efficient trials whereby treatments that promise much earlier are fast-tracked and less effective de-prioritized.

4.4. Real-time monitoring and predictive analytics

Clinical trials monitoring may leverage the power of AI since such real-time analysis of patient data will be useful in quickly establishing safety problems and adverse events. Besides, predictive analytics, powered by AI, will project the outcome from early patient data so that one can decide wisely. This ability is most important for the early identification of possible risks and benefits during a trial process for more effective and safer treatments.

5. The Perils of AI in Clinical Trials

Although the potential benefits of AI in clinical trials are huge, so are the challenges to be addressed for its ethical and effective application in this context. Figure 2 presents the major perils of AI clinical trials.

5.1. Data Privacy and Security Concern

The application of AI in clinical trials presupposes the collection and analysis of large amount of personal health data. This gives more rise to concerns of privacy and data security, particularly, of sensitive medical information. Ensuring AI systems are complaint with data protection regulation, such as General Data Protection Regulation in Europe is critically important. There is also the risk that AI systems might be vulnerable to cyberattacks, which could lead to a breach in patient confidentiality and the data being misused.

5.2. Algorithmic Bias and Fairness

AI algorithms are only as good as the data on which they have been trained. If the training data is biased or unrepresentative, it can mean that AI systems produce biased outcomes, affecting differentially certain groups of patients. This will become of major concern in clinical trials, where biased algorithms could permit unequal access to treatment or inaccurate predictions about treatment efficacy for different demographic groups. Ensuring AI systems are trained on diverse and representative datasets is critical to mitigating these risks.

5.3. Transparency and Explainability

Such concerns also exist with regard to the fact that most AI systems, especially those involving deep learning, are often "black boxes" for which it can be pretty difficult to understand how they came up with certain decisions or predictions. This may then cause problems during clinical trials, where maximum effort is made to understand the logic behind treatment decisions and the subsequent results as clearly as possible. In that light, it is incumbent on researchers and regulators to commence the process of developing AI systems that are not only accurate but also explainable, so that use in clinical trials is transparent and hence trustable.

5.4. Algorithm

The infusion of AI into clinical trials will obviously raise a new dimension of regulatory challenges since prevailing frameworks may not be fully ready to face the challenge of the complexity of AI-driven research. This will require up-to-date guidelines to be developed, taking into account the specificity connected with the AI applications, such as algorithm validation, data governance, and ethical concerns. As shown in Table 2, the integration of AI in clinical trials presents a range of promises, including enhanced precision and efficiency, alongside significant perils, such as ethical dilemmas and data privacy concerns, highlighting the need for careful consideration in its application. Also, a series of ethical questions arise from the use of AI in clinical trials about how automation might eventually substitute human judgment in areas critical to patient care and research. The challenge would be to balance the benefits of AI with the needs for human oversight and ethical accountability.

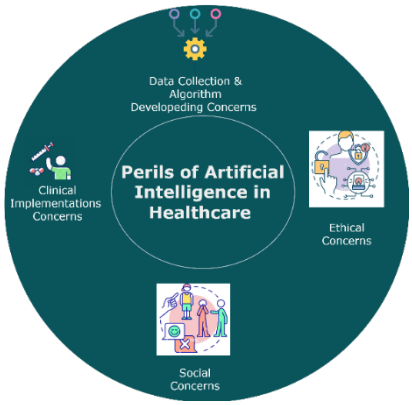


Figure 2: Perils of AI in Healthcare.

6. Recommendations for Future Research

1. Future research emphasis should be to enhance the interpretability and transparency of complex models, such as those dealing with deep learning, toward clinicians and regulators to instill trust in AI-driven decisions made during clinical trials.
2. Assess how AI integrates different types of data-imaging, genomics, clinical information-into one analysis that has the potential to advance personalized treatment strategies, and subsequently overall trial outcomes, for each data type.
3. Strive for models applying real-world evidence through artificial intelligence and wearables to EHRs and patient-reported outcomes to develop better generalizability and applicability of findings from clinical trials.
4. Design a framework directed towards the elimination of bias in artificial intelligence models in order to achieve equality regarding age, gender, and ethnicity. Further research works shall, therefore, be channeled towards the improvement of training datasets to make the AI assessment representative and inclusive.
5. The development of the AI framework should focus on international legislation requirements. Research is to be directed to make AI systems used in trials transparent and auditable, which again justifies that privacy and safety regulations from organizations such as the FDA and EMA are met.

7. Conclusion

Artificial intelligence might be very instrumental in changing clinical trials into medical research with the potential advantages of efficiency, accuracy, and scalability. Such benefits are hedged with major challenges that need to be navigated with care for

ensuring ethical and effective applications of AI. It is in the research community's best interest to address issues of data privacy, algorithm bias, transparency, and regulation to realize the power of AI in moving clinical trials forward without putting a patient's rights and well-being in jeopardy. In other words, AI in clinical research has a future if it is able to strike proper balance between innovation and ethical responsibility in ensuring that this technology gets used to its fullest for the betterment of all.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Askin, D. Burkhalter, G. Calado, S. El Dakrouni, Artificial Intelligence Applied to clinical trials: opportunities and challenges, *Health and Technology*, vol. 13, no. 2, 2023, doi:10.1007/s12553-023-00738-2.

[2] S. Piantadosi, Clinical trials: a methodologic perspective, John Wiley & Sons, 2024.

[3] H. Saeed, I. El Naqa, Artificial intelligence in clinical trials, Springer: 453–501, 2022.

[4] B.C. Stahl, Artificial Intelligence for a Better Future An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies, 2021.

[5] W.H. Organization, Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance, World Health Organization, 2024.

[6] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, *Nature Medicine*, vol. 25, no. 1, 2019, doi:10.1038/s41591-018-0300-7.

[7] N. Rieke, J. Hancox, W. Li, F. Milletarì, H.R. Roth, S. Albarqouni, S. Bakas, M.N. Galtier, B.A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R.M. Summers, A. Trask, D. Xu, M. Baust, M.J. Cardoso, "The future of digital health with federated learning," *Npj Digital Medicine*, vol. 3, no. 1, 2020, doi:10.1038/s41746-020-00323-1.

[8] M. Woo, "Trial by artificial intelligence," *Nature*, vol. 573, no. 7775, pp. S100–S102, 2019.

[9] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G.C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F.J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C.J. Kelly, D. King, J.R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J.J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K.C. Young, et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, 2020, doi:10.1038/s41586-019-1799-6.

[10] T. Dhar, N. Dey, S. Borra, R.S. Sherratt, "Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, 2023, doi:10.1109/tts.2023.3234203.

[11] S. Gerke, T. Minssen, G. Cohen, Ethical and legal challenges of artificial intelligence-driven healthcare, 2020, doi:10.1016/B978-0-12-818438-7.00012-5.

[12] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, 2017, doi:10.1038/nature21056.

[13] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: Past, present and future, *Stroke and Vascular Neurology*, vol. 2, no. 4, 2017, doi:10.1136/svn-2017-000101.

[14] T.M. Maddox, Clinical Decision Support in Statin Prescription - What We Can Learn from a Negative Outcome, *JAMA Cardiology*, vol. 6, no. 1, 2021, doi:10.1001/jamacardio.2020.4756.

[15] N. Fleming, "How artificial intelligence is changing drug discovery," *Nature*, vol. 557, no. 7707, 2018, doi:10.1038/d41586-018-05267-x.

[16] F.Y. Chiu, Y. Yen, Imaging biomarkers for clinical applications in neuro-oncology: current status and future perspectives, *Biomarker Research*, vol. 11, no. 1, 2023, doi:10.1186/s40364-023-00476-7.

[17] M. Yaqub, F. Jinchao, M.S. Zia, K. Arshid, K. Jia, Z.U. Rehman, A. Mehmood, "State-of-the-art CNN optimizer for brain tumor segmentation in magnetic resonance images," *Brain Sciences*, vol. 10, no. 7, 2020, doi:10.3390/brainsci10070427.

[18] K.D. Davis, N. Aghaeepour, A.H. Ahn, M.S. Angst, D. Borsook, A. Brenton, M.E. Burczynski, C. Crean, R. Edwards, B. Gaudilliere, G.W. Hergenroeder, M.J. Iadarola, S. Iyengar, Y. Jiang, J.T. Kong, S. Mackey, C.Y. Saab, C.N. Sang, J. Scholz, M. Segerdahl, I. Tracey, C. Veasley, J. Wang, T.D. Wager, A.D. Wasan, M.A. Pelleymounter, "Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities," *Nature Reviews Neurology*, vol. 16, no. 7, 2020, doi:10.1038/s41582-020-0362-2.

[19] A. Prelaj, V. Miskovic, M. Zanitti, F. Trovo, C. Genova, G. Viscardi, S.E. Rebuzzi, L. Mazzeo, L. Provenzano, S. Kosta, M. Favali, A. Spagnoletti, L. Castelo-Branco, J. Dolezal, A.T. Pearson, G. Lo Russo, C. Proto, M. Ganzinelli, C. Giani, E. Ambrosini, S. Turajlic, L. Au, M. Koopman, S. Delaloge, J.N. Kather, F. de Braud, M.C. Garassino, G. Pentheroudakis, C. Spencer, et al., Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review, *Annals of Oncology*, vol. 35, no. 1, 2024, doi:10.1016/j.annonc.2023.10.125.

[20] B. Rossi, B. Santos-Lima, E. Terrabuio, E. Zenaro, G. Constantin, Common Peripheral Immunity Mechanisms in Multiple Sclerosis and Alzheimer's Disease, *Frontiers in Immunology*, vol. 12, 2021, doi:10.3389/fimmu.2021.639369.

[21] E.T. Hébert, C.K. Ra, A.C. Alexander, A. Helt, R. Moisiuc, D.E. Kendzor, D.J. Vidrine, R.K. Funk-Lawler, M.S. Businelle, "A mobile just-in-time adaptive intervention for smoking cessation: Pilot randomized controlled trial," *Journal of Medical Internet Research*, vol. 22, no. 3, 2020, doi:10.2196/16907.

[22] S. Prabhu, K. Prasad, A. Robels-Kelly, X. Lu, AI-based carcinoma detection and classification using histopathological images: A systematic review, *Computers in Biology and Medicine*, vol. 142, 2022, doi:10.1016/j.compbiomed.2022.105209.

[23] W.M. van Oostveen, E.C.M. de Lange, "Imaging techniques in Alzheimer's disease: a review of applications in early diagnosis and longitudinal monitoring," *International Journal of Molecular Sciences*, vol. 22, no. 4, 2110, 2021.

[24] W. He, Z. Zhang, Y. Luo, R.T.K. Kwok, Z. Zhao, B.Z. Tang, Recent advances of aggregation-induced emission materials for fluorescence image-guided surgery, *Biomaterials*, vol. 288, 2022, doi:10.1016/j.biomaterials.2022.121709.

[25] A. Smart, P. Martin, M. Parker, Tailored medicine: Whom will it fit? The ethics of patient and disease stratification, *Bioethics*, vol. 18, no. 4, 2004, doi:10.1111/j.1467-8519.2004.00400.x.

[26] Z. Chaudhary, C. Lungren, A. Shah, "Tempus: Reshaping the Landscape of Cancer Care with Revolutionary Artificial

Intelligence," *Teens in Health AI in Healthcare Summer 2023 Journal Teens in Health Is a Teen Led Organization That Aims to Provide Open Access to Biological Research Skill Development through Researching and Writing Articles*, 148, 2023.

[27] K.H. Zou, C. Vigna, A. Talwai, R. Jain, A. Galaznik, M.L. Berger, J.Z. Li, The Next Horizon of Drug Development: External Control Arms and Innovative Tools to Enrich Clinical Trial Data, *Therapeutic Innovation and Regulatory Science*, vol. 58, no. 3, 2024, doi:10.1007/s43441-024-00627-4.

[28] D. Hartl, V. de Luca, A. Kostikova, J. Laramie, S. Kennedy, E. Ferrero, R. Siegel, M. Fink, S. Ahmed, J. Millholland, A. Schuhmacher, M. Hinder, L. Piali, A. Roth, Translational precision medicine: an industry perspective, *Journal of Translational Medicine*, vol. 19, no. 1, 2021, doi:10.1186/s12967-021-02910-6.

[29] M. Shanmuga Sundari, H.R. Penthala, A. Mogullapalli, M.M. Ammangatambu, "AI-Based Personalized Drug Treatment," *Artificial Intelligence and Machine Learning in Drug Design and Development*, 369–406, 2024.

[30] J. Geraci, R. Bhargava, B. Qorri, P. Leonchyk, D. Cook, M. Cook, F. Sie, L. Pani, "Machine learning hypothesis-generation for patient stratification and target discovery in rare disease: our experience with Open Science in ALS," *Frontiers in Computational Neuroscience*, vol. 17, 2023, doi:10.3389/fncom.2023.1199736.

[31] S.M. Williamson, V. Prybutok, Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare, *Applied Sciences (Switzerland)*, vol. 14, no. 2, 2024, doi:10.3390/app14020675.

[32] R. Schwartz, R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, P. Hall, Towards a standard for identifying and managing bias in artificial intelligence, US Department of Commerce, National Institute of Standards and Technology, 2022.

[33] N. Hendrix, D.L. Veenstra, M. Cheng, N.C. Anderson, S. Verguet, "Assessing the Economic Value of Clinical Artificial Intelligence: Challenges and Opportunities," *Value in Health*, vol. 25, no. 3, 2022, doi:10.1016/j.jval.2021.08.015.

[34] A. Jungbauer, G. Ferreira, M. Butler, S. D'Costa, K. Brower, A. Rayat, R. Willson, "Status and future developments for downstream processing of biological products: Perspectives from the Recovery XIX yield roundtable discussions," *Biotechnology and Bioengineering*, 2024.

[35] P.G.R. de Almeida, C.D. dos Santos, J.S. Farias, "Artificial Intelligence Regulation: a framework for governance," *Ethics and Information Technology*, vol. 23, no. 3, 2021, doi:10.1007/s10676-021-09593-z.

[36] A. Nassar, M. Kamal, "Ethical Dilemmas in AI-Powered Decision-Making: A Deep Dive into Big Data-Driven Ethical Considerations," *International Journal of Responsible Artificial Intelligence*, vol. 11, no. 8, 2021.

[37] E. Nasarian, R. Alizadehsani, U.R. Acharya, K.-L. Tsui, "Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework," *Information Fusion*, 102412, 2024.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by- sa/4.0/).



Muhammad Yaqub working as postdoctoral researcher at Hunan University. He received his Bachelor's and Master of Science in Computer Science from COMSATS University Islamabad, Pakistan, a top-ranked institution for IT education in Pakistan. He earned his Ph.D. in Electronic Science and Technology from Beijing University of Technology, China, in 2023, where he made significant contributions to the development of innovative deep learning-based schemes for medical image analysis, including segmentation, classification, and image reconstruction. With a strong academic background and a track record of over 30 publications in reputable international journals, Dr. M Yaqub is renowned for his expertise in applying cutting-edge technologies to advance cancer and oncology research. His work has consistently demonstrated a commitment to excellence in research and a keen analytical mindset.



Lan He working as associate professor at Hunan university, Changsha, China. She graduated from the School of Life Sciences of Northeast Normal University in 2010 and was later recommended to the Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, where she received her Ph.D. in Science in 2016. In the same year, she entered Shanghai Tech University for postdoctoral research. She has published many papers as the first author or corresponding author in international authoritative journals, including Advanced Science, Cellular & Molecular Immunology, PNAS, etc. She has participated in the writing of two authoritative immunology books and one English book for Springer magazine. She is currently undertaking one National Natural Science Youth Fund, one Hunan Provincial Natural Science Youth Fund, and one Changsha Natural Science Fund.

Received: 07 August 2024, Revised: 14 September 2024, Accepted: 17 September 2024, Online: 22 September 2024

DOI: https://dx.doi.org/10.55708/js0309003

Fingerprint Bio-metric: Confronting Challenges, Embracing Evolution, and Extending Utility - A Review

Diptadip Maiti[*,1], **Madhuchhanda Basak**[2], **Debashis Das**[3]

[1]Department of CSE, Techno India University, West Bengal, 700091, India

[2] Department of CSE, Brainware University, West Bengal, 700125, India

[3] Department of CSE, Dr. Sudhir Chandra Sur Institute of Technology & Sports Complex, West Bengal, 700074, India

*Corresponding author: Diptadip Maiti, Techno India University, West Bengal, 700091, India, Contact No & Email: +919433736910, diptadipmaiti@gmail.com

ABSTRACT: As documented in recent research, this review offers a thorough examination of the intricate subject of fingerprint authentication, including a wide range of issues and applications. Addressing problems like non-linear deformations and enhancing picture quality, which are frequently reduced by sophisticated improvement and alignment techniques are important components of fingerprint image authentication. Countering security concerns such as spoofing is a major focus of Automated Fingerprint Identification Systems and necessitates the use of sophisticated cryptographic techniques and liveness detection. In order to accomplish speedier identification processes, the paper emphasizes the advancements made in fingerprint indexing and retrieval, with a focus on deep learning technologies and minutiae-based methodologies. Furthermore, fingerprint authentication is used for a variety of age groups, including neonates, where it is essential for identification verification and the management of medical records. The paper also highlights the wider uses of fingerprint technology, such as improved crime detection skills, insights into age-related features, and contributions to medical diagnostics. This review provides a thorough overview of the latest developments and potential future directions in fingerprint authentication by combining state-of-the-art methodologies and analysing technical details, implementation challenges, and security issues. This captures the dynamic and important role of this biometric technology.

KEYWORDS: Biometric Authentication, Fingerprint Identification System, Biometric Security, Biometric Application

1. Introduction

In the rapidly evolving landscape of biometric technology, fingerprint authentication has emerged as a cornerstone of identity verification and security systems. The intricate and unique ridge patterns present on human fingertips provide a reliable and convenient means of confirming individuals' identities. As societies transition towards digitalization and interconnectedness, the role of fingerprint biometrics becomes increasingly crucial in ensuring secure access to systems, facilities, and personal information. This paper delves into the intricate web of fingerprint biometric research, offering a comprehensive overview of the challenges faced, innovative solutions devised, and emerging trends that collectively shape the present and future of this dynamic field. Figure 1 depicts a few fingerprint images that have been collected through different technologies. Figure 1(a) represents fingerprint collected by ink and paper method, Figure 1(b) shows fingerprint collected by a digital scanner and Figure 1(c) shows fingerprint collected from a crime scene.

We will go through a number of fingerprint biometric dimensions in this exploration, each with its own opportunities and limitations. The challenge of authenticating unformatted fingerprint photos is significant because of non-linear deformations, different pressures during acquisition, and a range of environmental factors. [1]. In response, researchers have developed intricate algorithms harnessing machine learning, convolutional networks, and auto encoders to enhance image quality and improve matching accuracy. Security vulnerabilities inherent in Automated Fingerprint Identification Systems (AFIS) call for advanced strategies to safeguard against threats such as spoofing and data tampering. Biometric cryptosystems and cancellable templates are among the innovative solutions that fuse cryptographic techniques with biometric data, forging new frontiers in secure authentication [2]. The realm of fast fingerprint indexing and searching is another pivotal arena, demanding the fusion of speed and precision in matching large datasets. Leveraging fingerprint features and deep learning algorithms, researchers have crafted strategies to expedite the matching process while ensuring accurate results. Furthermore, the application of fingerprint biometrics extends beyond traditional identity verification. From diagnosing medical conditions through changes in fingerprint patterns to detecting drug consumption and even determining handedness, the fingerprint's unique attributes are being harnessed for a diverse array of purposes. As this exploration unfolds, it underscores not only the strides made

in fingerprint biometrics but also the ethical considerations and societal implications that accompany the growing integration of biometric data into various aspects of modern life. As researchers continue to innovate and push the boundaries of what is possible, fingerprint biometrics remain a cornerstone in the pursuit of a secure and interconnected future [3].

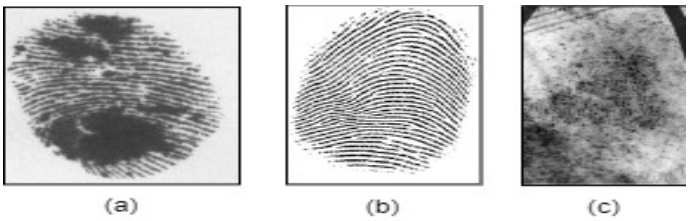


Figure 1: Different Source FP Images: (a) Rolled ink (b) Live scan (c) Latent

The objective of this review is to comprehensively explore the landscape of fingerprint authentication as documented in previous literature. It aims to delve into the technical challenges encountered in fingerprint image authentication, including addressing non-linear deformations, enhancing image quality, and implementing alignment methods. The review also seeks to examine the security considerations within Automated Fingerprint Identification Systems (AFIS), focusing on strategies to combat spoofing through liveness detection and advanced cryptographic measures. Additionally, it aims to analyse the methodologies and advancements in rapid fingerprint indexing and retrieval, particularly emphasizing minutiae-based approaches and the integration of deep learning techniques. Furthermore, the review intends to assess the diverse applications of fingerprint authentication, from identity verification in infants to its implications for medical history tracking, as well as its broader roles in medical diagnostics, age-related studies, handedness detection, and crime prevention. Through a thematic analysis of existing literature, this review aims to provide insights into the current state-of-the-art methodologies, implementation challenges, and security concerns within the field of fingerprint authentication.

To present a thorough overview of developments in the field of biometric authentication, we carefully reviewed publications published in the past ten years as part of the data collecting and analysis for this review study. Each study was methodically identified and categorized according to its purported benefits and drawbacks, accuracy measures, and procedures used. To comprehend these research' methods to algorithmic processing, performance evaluation, and biometric data gathering, we carefully examined their techniques. The study aimed to evaluate the efficacy and accuracy rates of various biometric approaches, taking into account the advantages and disadvantages of each method. Through the consolidation of this data, we were able to identify patterns and breakthroughs, evaluate the development of biometric technologies, and provide perspectives on the condition of the technology sector at large. This comprehensive study seeks to offer a fair assessment of the advancements made, point out areas in need of development, and determine future paths for biometric authentication research.

In this study, we attempt to investigate various state-of-the-art developments in finger print biometric practical application. The format of this essay is as follows: The difficulties and security risks associated with latent fingerprint authentication and automatic fingerprint authentication systems are discussed in Section 2. The method for all-purpose fingerprint user identification is described in Section 3. Fast fingerprint indexing and searching are covered in Section 4. The difficulties and solutions for analysing and identifying infants' fingerprints are described in Section 5. Multi-modal biometric techniques are discussed in Section 6. Comparative comparison of various approaches is the main topic of Section 7. Section 8 concludes with some final thoughts.

2. Challenges in Authentication

2.1. Authentication of Unformatted Fingerprint Image

In the field of fingerprint authentication, raw fingerprint images can be obtained from different sources, such as rolled ink prints, live scans, and latent prints. These images often suffer from non-linear deformations and poor quality, making authentication challenging. Non-linear deformations can arise due to variations in finger pressure and improper finger placement during image acquisition. These deformations can lead to unsatisfactory matching scores and impact the accuracy of fingerprint authentication systems. Low-quality fingerprint images can also result from factors like image scanner noise, partial prints collected from crime scenes (latent prints), and various skin conditions. These low-quality prints can be categorized as dry, wet, damaged, dotted, and blurred. Dry prints have minimal contact with the scanning surface, while wet prints result in smudged ridges and valleys due to extensive contact. Damaged prints may arise from scars or skin issues, while dotted prints are caused by excessive sweating. Blurred prints occur when motion or unclear ridges during scanning lead to image blurriness. To address these challenges, various techniques have been developed for fingerprint enhancement and alignment. Traditional methods such as Gabor filtering and adaptive boosted spectral filtering are commonly used for enhancing fingerprint images. Alignment techniques and cluster-based methods aim to rectify non-linear deformations, improving the accuracy of matching in automated fingerprint identification systems (AFIS). Figure 2 shows fingerprint feature extraction using convolution neural network, where three parallel CNN extract features from the fingerprint image and after the merging of the three features to form a unique feature descriptor, which is passed to a classifier for classification task.

Because of their high dimensionality, minutiae descriptors, which are widely used in fingerprint identification algorithms, face considerable hurdles. These descriptors are robust for fingerprint pattern identification because they capture fine features. However, the noise and volatility present in real-world settings might have a significant impact on their performance. Noise may cause mismatches during matching procedures by distorting minute points. This noise can originate from various sources, such as defective sensors or low-quality images. Accurate identification is made more difficult by unpredictable variations in fingerprint traits, such as ageing or changes brought on by injuries. Thus, even though minutiae descriptors are excellent at

collecting distinctive fingerprint features, their effectiveness depends on how well noise and unpredictability are handled to provide safe and dependable biometric identification systems.

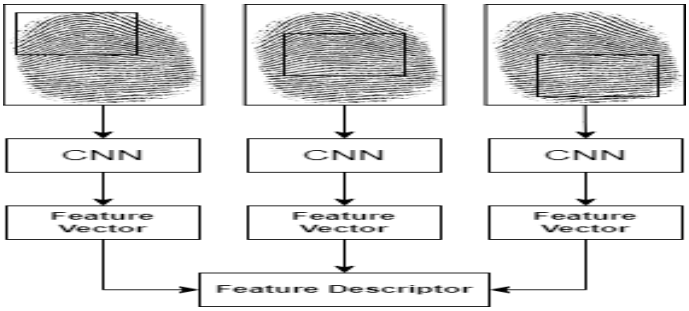


Figure 2: Extraction of Minutiae Descriptor using CNN

A novel latent fingerprint matching approach is presented by [4], which makes use of a descriptor-based Hough transform for latent prints together with a robust alignment technique. Their approach achieved a noteworthy rank-1 accuracy of 53.5% on NIST SD27, outperforming two commercial matchers and a non-commercial algorithm. Singular point-based alignment is a problem, as performance depends on the quantity of minutiae and the overall print quality. Refining features with top-down information from an exemplar print, in [5], the authors presented a feedback system for latent fingerprint matching. When combined with a cutting-edge matcher, it improved identification accuracy in the NIST SD27 and WVU databases by 0.5–3.5%. Finding a balance between the advantages of feedback and the complexity of the system, as well as the example print quality, is a challenge. Deformable Minutiae Clustering was proposed by [6] for latent fingerprint recognition, improving minutiae matching via cluster merging and Thin Plate Spline modelling. It outperformed previous techniques with up to 85.6% (Cylinder-Codes) and 83.3% (m-triplets) accuracy across different fingerprint databases. Scalability issues for latent-to-latent identification and algorithm speed are among the limitations. Using feature selection and random decision forest classification, the authors in [7] provided an adaptive latent fingerprint segmentation approach that achieved state-of-the-art performance on three databases. In order to balance accuracy and selection time, their method includes a unique SIVV-based metric for segmentation evaluation and modified RELIEF feature selection. Constraints about the applicability of the SIVV metric and assumptions about consistent ground truth are examples of limitations. A collaborative filtering paradigm for fingerprint enhancement was proposed by [8]. It involves pre-enhancing using Gabor filters and spatial patch-based enhancement utilizing spectral diffusion after that. Test findings on FVC2004 datasets showed that this approach was better than other approaches such as the Gabor filter and VeriFinger Algorithm Demo. The study is limited by fixed patch size effects and sensitivity to input fingerprint quality, which calls for additional investigation to fully evaluate resilience. FingerNet, a CNN-based method for latent fingerprint enhancement, was introduced by [9]. It has shared convolution and deconvolution layers for orientation guiding and noise reduction. It demonstrated potential for improvement with larger datasets and ground truth ROI annotations, as demonstrated by its top-1 matching accuracy of 47.7% when tested on NIST SD27 using data collected from SD4. Using Convolutional Neural Networks (ConvNets) for ridge flow estimation and minutia descriptor extraction. In [10], the authors proposed an automated latent fingerprint recognition system that achieved rank 1 identification accuracies of 64.7% and 75.3% on NIST SD27 and WVU latent databases. Poor ridge quality, background noise, a tiny friction ridge area, reliance on manual ROI selection, long processing times, and scaling issues are some of the limitations. A technique to improve Automated Fingerprint Identification Systems (AFIS) by utilizing uncommon minutiae was suggested by [11]. This involves altering similarity scores that are based on least squares fitting mistakes in order to improve matching accuracy. It greatly improves the rank identification accuracy of minutiae-based matchers, as demonstrated by tests conducted on the GCDB forensic database. Future study on generalization and database size effects is suggested by the challenges of dataset restrictions and manual intervention for latent AFIS. An end-to-end latent fingerprint recognition system with automated ROI cropping, minutiae extraction based on deep learning, and texture template creation was promoted by [12]. Tested against a background of rolled prints on many datasets, yielding rank-1 retrieval rates varying from 7.6% to 69.4%. Among the drawbacks are difficulties in identifying fine details on low-quality photos, trimming dry latents, and the requirement for more varied operating databases for thorough training. By utilizing fine-coarse parallelism and asynchronous processing for latent fingerprint recognition. In [13], the authors invented ALFI and achieved a 22x speed gain over state-of-the-art approaches with equivalent accuracy. Tested on NIST SD27, they report accuracy using Equal Error Rate and F1-score, recognizing dataset size constraints, restricted classification strategies, and unknown uses beyond fingerprints, such as DNA analysis. LQMetric, an automated method for evaluating latent fingerprint quality and forecasting the probability of a rank-1 hit in the FBI's NGI AFIS system, was suggested by [14]. It produced a 61.4% match with human examiner assessments using local clarity maps and image analysis tools; it worked well for NGI searches but was not uniformly applicable to other AFIS methods. For improving latent fingerprint images. In [15], the authors provide a progressive GAN-based technique that consists of off-line training and repeated on-line testing phases. Based on CMC curve metrics, evaluation on the NIST SD27 dataset demonstrates better performance than previous models. Constraints on dataset quantity, processing cost, and possible inefficiency with subtle ridge features in latent prints are some of the limitations. An automated latent fingerprint identification system using DCNN-FFT augmentation for minutiae extraction and matching was described by [16]. It achieved 100% rank-1 identification on the FVC2002 and FVC2004 databases and 84.5% on the NIST SD27. The results show increased recall, F1 scores, and precision; however, computational time and dataset diversity are acknowledged as constraints. Using synthetic data for training. In [17], the author provide a deep nested UNets architecture for automated latent fingerprint segmentation and improvement. It performs better than current methods in fingerprint recognition and segmentation accuracy when tested on the NIST SD27 and IIITD-MOLF

databases. One of the main drawbacks is the lack of high-quality and low-quality coupled fingerprint picture pairs in the current databases. This can be overcome by creating synthetic latent images for training. Using dense fingerprint patch alignment and matching. In [18], the authors promoted a non-minutia latent fingerprint registration method that outperformed previous techniques in registration when tested on the NIST27 and MOLF datasets. Its drawbacks include its dependence on 2D data, higher processing requirements, and possible inadequacy for poor-quality picture data. Using an algorithm based on Lindeberg's scale selection method. In [19], the authors revealed integrating pores (level 3 characteristics) alongside minutiae for latent fingerprint identification. When combined with pores, as opposed to just minutiae matching, greatly improves recognition rates on the IIITD Latent database. A small dataset size, a focus plane restriction for pore extraction, and computational costs that affect matching time are among the limitations. For latent fingerprint identification. In [20], the authors proposed the "Ratio of Minutiae Triangles" (RMT) algorithm, which makes use of local minutiae arrangements without pre-alignment. tested with Rank-1 recognition accuracy of 80% and 63.8% on the FVC2004 and NIST SD27 databases, respectively. Poor quality latent cases and the requirement for better feature extraction and matching phase integration for large databases are acknowledged constraints. In order to achieve accurate latent fingerprint enhancement and segmentation. In [21], the researchers proposed a hybrid model that combines the Chan-Vese approach for segmentation and Edge Adaptive Directional Total Variation (EDTV) for enhancement. Accurate Rank-1 identification was achieved with 61.24% and 70.16% on NIST SD27 and WVU DB databases, respectively. Restrictions include the size of the database and the requirement for additional testing on intricate latent fingerprints. Using Chan-Vese, an adaptive latent fingerprint segmentation and matching strategy based on the EDTV model was given by [22]. On the NIST SD27 dataset, this approach achieved AUCs of 72% and 70.59% for ROC and CMC curves, respectively. Limitation includes poor accuracy of existing techniques for latent fingerprint segmentation. Using scale and rotation invariant minutiae characteristics. In [23], the authors proposed an automatic latent fingerprint identification system that significantly improved Rank-1 identification accuracy on NIST SD27 and FVC2004 datasets. On NIST SD27, CLMP-NRS produced the maximum Rank-1 accuracy of 93.80%. Handling incomplete fingerprints with limited minutiae characteristics and relying on the quantity of retrieved minutiae for matching performance are two limitations. MinNet, a minutia patch embedding network for latent fingerprint identification that optimizes spatial and angular minutiae distribution, is introduced by [24]. Reaching cutting-edge outcomes, assessed on many public and private datasets, such as FVC-Latent and NIST SD27, with rank-1 accuracies of 85.88% and 92.39%, respectively. Its limitations include the inability to handle fingerprints that are partially or substantially deformed and its reliance on high-quality latent pictures for precise minutiae extraction. To achieve state-of-the-art accuracy and efficiency in latent fingerprint indexing, the authors in [25] offer a multi-scale fixed-length representation technique. performed better than other indexing techniques like PDC and DeepPrint when tested on a variety of datasets, including Hisign and NIST SD27. possibility performance variance on simulated fingerprints and scope-limited optimization are among the limitations, indicating possibility for more improvements in feature representation. Using a residual encoder-decoder architecture and frequency-domain loss function optimization, the authors in [26] proposed a deep learning model for latent fingerprint enhancement. superior than current methods in terms of rank-25 and rank-50 accuracy when evaluated using the IIIT-Delhi MOLF database. The ridge-based approach of the technology and the requirement for additional refinement in deleting undesired portions of images and testing on a variety of databases are its limitations. With an emphasis on fairness in prediction and decision-making, the authors in [27] provided a technique to and reduce biases in automated algorithms working with latent fingerprint pictures. They evaluate the effect on automatic matching of latent fingerprints using covariate-specific ROC curves produced from regression models taking quality and demographics into account. Results demonstrate that, compared to assessments without quality, the suggested covariate-adjusted ROC curves conditioned on image quality and demographics offer a more informative assessment. The quality measurement algorithm's underlying assumptions and dataset restrictions are examples of limitations. It is believed that more research on a variety of datasets is crucial for method validation. ULPrint, a Universal Latent Fingerprint Enhancer, was introduced by [28], who used Mix Visual Transformer (MiT) SegFormer-B5 encoder architecture and Ridge Segmentation with UNet. By using directed blending of predicted ridge masks, the technique improves latent fingerprints and addresses issues with a variety of latent types. Tests conducted on both synthetic and actual datasets show notable gains in accuracy; nonetheless, there are still issues such as annotator subjectivity and limited databases. FingerGAN, a GAN-based technique for latent fingerprint enhancement, was presented by [29]. It optimizes minute information through adversarial, perceptual, and reconstruction losses. Tested on NIST SD27 and IIIT-Delhi MOLF, FingerGAN outperforms state-of-the-art techniques, however it has drawbacks such as computational complexity and dependence on high-quality rolled fingerprints for data production. For latent fingerprint recognition, the authors in [30] suggest a hybrid method integrating local (minutiae and virtual minutiae) and global features (AFR-Net embeddings). With a multi-stage matching paradigm and Squeeze U-Net CNN for augmentation, they attain an average rank-1 retrieval rate of 71.22% on multiple datasets. Failure instances including severe rotations and overlapping patterns are noted, as are challenges such as low contrast and occlusion. ACSACO, a hybrid approach combining Cuckoo Search and Ant Colony Optimization for latent fingerprint identification, was proposed by [31]. ACSACO beats individual algorithms when tested on the NIST SD-27 dataset, achieving excellent precision and recall for prints of good quality but lower accuracy for prints of poor or ugly quality. Complex backdrop problems and overlapping prints are among the limitations, indicating the need for additional optimization strategies and wider dataset validation. In order to enhance DeepPrint for recognition systems, in [32], the authors developed a CycleGAN-based technique to create artificial latent fingerprints. Their ap-

proach improves TDR at 0.01% FAR from 23.64% (MSP) to 45.45% (synthetic) when tested on NIST SD27. Reliance on a single matcher and the lack of functional latent databases are among the limitations.

Latent fingerprint identification is the painstaking process of locating and examining fingerprints from crime scenes, frequently with the aid of cutting-edge methods to improve their contrast and visibility. New levels of security are introduced when these latent prints are scanned and input into automatic fingerprint identification systems (AFIS). This meticulous manual process is essential for correct identification. Although AFIS significantly improves fingerprint matchings scalability and efficiency, it also brings with it potential vulnerabilities that were not as noticeable in manual systems. Problems like algorithmic flaws, spoofing attempts, and data breaches emphasize the necessity of strict security protocols to preserve the accuracy of fingerprint data. In order to guarantee that the improvements in fingerprint identification technology do not jeopardize its dependability and efficacy, it is crucial to comprehend the shift from manual latent fingerprint analysis to automated methods.

2.2. Security Threats in AFIS

Digital user authentication through fingerprint biometrics relies on Automated Fingerprint Identification Systems (AFIS) to verify individuals' identities. The concept of AFIS was initially developed by the US Federal Bureau of Investigation (FBI). The AFIS process can be divided into two phases: Enrolment and Identification. In the Enrolment phase, users' fingerprints are registered in the system. The Identification phase matches query fingerprints against stored templates for authentication. During Enrolment, fingerprint images undergo preprocessing, feature extraction, and template creation. In the Identification phase, query fingerprints are processed and matched against stored templates to identify valid users. Securing a biometric system is crucial due to the non-changeable nature of biometric traits. An AFIS needs robust security measures to safeguard user information from potential attackers. Various components within an AFIS can be exploited by attackers. These attack points include potential vulnerabilities in the biometric data acquisition process, template extraction, and communication channels. Several known attacks can compromise the security of an AFIS, including spoofing (presenting fake biometric data), exploiting similarity, zero-effort attempts (using attacker's own biometric to impersonate a legitimate user), physical destruction of the biometric sensor, replay attacks (intercepting and replaying biometric signals), communication channel attacks (cutting or altering channels), and more. These attacks can lead to unauthorized access, denial of service, and other security breaches.

2.2.1. Liveness Detection

To address these security challenges, researchers have developed preventive measures and countermeasures, such as liveness detection to determine if a fingerprint is from a living person or a spoof. Various methods have been proposed to detect spoof attacks that use materials like wood, glue, and gelatine to fabricate fingerprint spoofs. Figure 3 shows the process of liveness detection using auto encoder where live and spoof fingerprint images are passed to a encoder to learnt two different latent representation of the image and then it is passed to decoder to reconstruct the live and spoof fingerprint, this way the architecture knows the difference between a spoof and live image and can detect them.

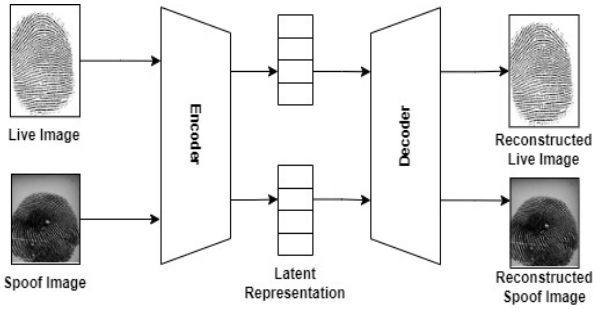


Figure 3: Process of Fingerprint Spoof Detection

In order to improve the resilience of biometric authentication systems against spoofing attacks, the authors in [33] present an approach that combines SURF, PHOG, and Gabor wavelets with low-level characteristics and shape analysis to detect liveness in fingerprint images. Achieving an average EER of 3.95%, their algorithm, which uses PCA for dimensionality reduction and various classifiers, such as SVM and Random Trees—beams above the previous record of 9.625%, validated across several databases, including LivDet 2011 and LivDet 2013. However, more testing on a range of sensors and materials is acknowledged for a thorough robustness assessment. A software-based fingerprint liveness detection technique is presented by [34]. It uses image gradient co-occurrence arrays for feature extraction and SVM classification. It attains greater accuracy on LivDet09DB and LivDet11DB datasets than state-of-the-art methods, in spite of drawbacks like quantization-induced information loss and high-dimensional feature vectors derived from higher-order co-occurrence arrays. Convolutional neural networks could improve gradient measurement; however, dimensionality reduction methods could be needed to handle higher-order arrays. In order to achieve state-of-the-art accuracy across LivDet datasets, the authors in [35] proposed a CNN-based technique for fingerprint spoof attack detection that makes use of local patches centred on minutiae for fine-grained analysis. The approach shows effectiveness against different spoof materials and testing settings, with an average accuracy of 99.03% on LivDet 2015, beating the 95.51% of competition winners. The need for more diverse datasets and ethical concerns about the use of biometric data are two acknowledged constraints, though. A fingerprint liveness detection technique using a BP neural network and difference co-occurrence matrices for improved textural features was suggested by [36]. The system builds input data based on these matrices and uses pre-trained networks to forecast classification accuracy, achieving higher detection accuracy on the LivDet 2013 database over earlier techniques. One of the limitations is the lack of diagonal direction difference co-occurrence matrices, which could more accurately represent the properties of the image. For real-time fingerprint liveness detection, the authors in [37] promoted a semi-supervised stacked autoencoder-based method that substitutes learnt representations for hand-

designed features. Exhibiting efficaciousness on the LivDet 2011 and 2013 datasets, the approach attains satisfactory performance, but with a recognition of constraints including dataset magnitude, model initialization, and difficulties with substandard quality, excessive brightness, or noisy images. Adversarial attacks on deep neural network-based fingerprint liveness detection were described by [38], exposing flaws in cutting-edge models. By applying FGSM, MI-FGSM, and Deepfool techniques, they illustrated how models might mistakenly identify phony fingerprints as real ones with minute variations. The evaluation of the LivDet 2013 and LivDet 2015 datasets made clear that deep learning applications require more reliable detection models and anti-adversarial techniques. With a score-level fusion approach that combines liveness detection and fingerprint matching, the authors in [39] took first place with 96.88% accuracy in the Fingerprint Liveness Detection Competition 2019. It was evaluated using the LivDet2015 and LivDet2019 datasets, taking into account restrictions on finger pressure and device time that may effect detection accuracy. It used ONNS for similarity, Slim-ResCNN for FLD score, and LR classifiers for fusion. A DenseNet optimized with a genetic algorithm was provided by [40] for fingerprint liveness detection, and it achieved 98.22% accuracy on a mixed Livdet dataset. Despite constraints in dataset size and computational cost for real-time applications, the method outperforms state-of-the-art efforts on LivDet 2009, 2011, 2013, and 2015 datasets by utilizing ROI extraction and specialized mutation operators. A liveness identification approach utilizing the Circular Gabor Wavelet (CGW) algorithm and SVM was presented by [41], which achieved 99.968% accuracy in differentiating between real and false fingerprints. The approach, which has been tested on 272 samples from optical and capacitive sensors, yields encouraging results but still needs to be assessed further against spoofing techniques used in the real world. A multi-modal liveness detection method incorporating fingerprint and iris inputs, leveraging statistical texture features and spatial analysis, was advocated by [42]. The approach outperforms sum-rule and product-based approaches with high precision (94.7%) for fingerprint detection and 97.8% accuracy for decision-level fusion. recognizing its limits, such as its reliance on dataset size and its usefulness against particular attack types. Using the LivDet 2013 and 2015 datasets, the authors in [43] presented FLDNet, a lightweight CNN architecture for fingerprint liveness detection that achieves state-of-the-art performance. FLDNet solves problems such as accuracy on tiny size fingerprints and robustness against unknown spoof materials, and achieves 1.76% Average Classification Error (ACE) over all sensors with a redesigned dense block and attention pooling layer. A one-class convolutional auto encoder for fingerprint presentation attack detection is proposed by [44], and it achieves a D-EER of 2.00% on a dataset of 24,050 fingerprint images. Among the method's drawbacks include its inability to generalize to additional modalities, the requirement for distinct models for various attack substrates, and uncertainties about robustness and adversarial attacks. A weighted multi-modal CNN-based FLD technique was introduced by [45], which used ROI operation and feature fusion for improved performance. It achieves excellent accuracy in different evaluation circumstances, outperforming current approaches on LivDet 2011, 2013, 2015, and NUAA datasets. Two drawbacks are the need for extensive training datasets and performance variability brought on by sensor diversity. EaZy learning, an adaptive ensemble learning model for fingerprint liveness detection, was suggested by [46]. On LivDet 2011, 2013, and 2015 datasets, it achieved average accuracies of 60.49% and 67.80%. It solves drawbacks such as reliance on clustering techniques and dataset size, as well as potential performance concerns in unknown contexts due to lack of variety in training data by clustering training data and integrating predictions using weighted majority voting. A novel approach using transformers and GANs to improve fingerprint presentation attack detection (PAD) generalization across sensors and materials was developed by [47]. With LivDet2015, the approach achieves an accuracy gain from 68.52% to 83.12%, addressing the limits of accuracy degradation in cross-sensor situations and inadequate generalization. Using a "transient liveness factor" (TLF) and picture quality measures, the authors in [48] suggest a person-specific FPAD technique that achieves 100% accuracy in spoof presentation detection. They admit constraints such as the short dataset size and the difficulty of generalizing across different materials while conducting experiments with a dataset of 30 live photos and 138 spoof samples from various attackers. In comparison to previous methods, the authors in [49] proposed a Fingerprint Liveness Detection (FLD) technique that improved accuracy by 1.0% on average by integrating AlexNet, VGG16, and ResNet CNNs with a genetic algorithm for feature weighting. Enhancing detection performance on several Livedet datasets, the strategy addresses the drawbacks of fixed-scale inputs and possible challenges in training the model due to over-abundance of features. A CNN-based fingerprint liveness detection technique was suggested by [50], which outperformed SVM and CNN+SVM hybrid techniques on the LivDet2015 dataset. Measured by precision, specificity, F1 score, and accuracy metrics, preprocessing approaches and feature extraction methods improve classification accuracy. Given that different materials provide different obstacles when it comes to creating false fingerprints, more study is necessary to improve classification performance in these scenarios. A multi-filter framework for fingerprint liveness detection utilizing hand-crafted features was promoted by [51], who achieved improved performance on LivDet test datasets. The method achieves an average accuracy of 99.15% and an average classification error of 0.85% on the LivDet 2015 dataset, outperforming state-of-the-art techniques. It involves procedures such as data augmentation, preprocessing, feature fusion, and dimensionality reduction. The authors recognize that lengthy processing times and complex parameter setting are real usability barriers. In order to counteract forced and phony fingerprint attacks, the authors in [52] introduced MFAS, a Micro-Behavioural Fingerprint Analysis System that records fingertip micro-behaviour throughout time. After being tested on 24 subjects, MFAS achieves 100% accuracy in identifying voluntary versus forced fingerprint placements, as well as 100% true positives. Though encouraging, the system can run into problems in realistic situations, requiring more study to improve technology and comprehend its application. The incremental learning model A-iLearn, proposed by [53], addresses the stability-plasticity conundrum in spoof fingerprint detection by adaptively integrating base

classifiers. A-iLearn increases accuracy on new fake materials by up to 49.57% without appreciable stability loss when compared to baseline models. It is tested on LivDet 2011, 2013, and 2015 datasets using a variety of spoof materials and sensors. Some drawbacks include the possibility of over fitting and the requirement for more research into feature integration techniques. By combining texture and sweating pore characteristics, the authors in [54] provided a static-based fingerprint liveness detection technique. An auto encoder minimizes feature dimensionality for binary classification using a softmax classifier by quantifying textural properties and pore activity. Testing the approach on LivDet 2013 and LivDet 2015, with an ACE of 0.11-0.24%, it recognizes the difficulties in maintaining several algorithms for hardware-based detection while addressing restricted pore feature-based techniques. A lightweight fingerprint liveness detection network called LFLDNet is proposed by [55]. It makes use of style transfer, foreground extraction, and an enhanced ResNet with MHSA. 95.27% accuracy on small-area fingerprints and 1.72% average classification error across sensors were achieved during evaluation on the LivDet2011, LivDet2013, and LivDet2015 datasets. The authors intend to investigate improved FLD technologies and GAN-based models for cross-sensor generalization against fingerprint deception assaults, while acknowledging the possibility of speed variations. Using CNNs and adversarial data augmentation, the authors in [56] proposed an FPAD technique and won the LivDet2021 competition. The method uses clean and adversarial fingerprints for a three-stage training process that yields good accuracy metrics and an EER of 0.036. The size restrictions of the dataset and possible susceptibilities to adversarial assaults are acknowledged by the authors. Using LPDJH image descriptors and deep learning, the authors in [57] introduced a fingerprint liveness detection technique that achieved good accuracy across LivDet datasets. Their approach shows competitive results on the LivDet 2011 dataset, with an average EER of 3.95%. Because there aren't enough test samples, the authors point out possible difficulties in identifying low-resolution fingerprints and calculating accuracy for particular sensors. A CNN-based fingerprint liveness detection technique was introduced by [58] using the Socofing dataset. The model performed well, achieving 98.964% accuracy with a FAR of 0.215% and a FRR of 7.251%. Computational limitations, inconsistent fingerprint quality, and unequal dataset distribution are some of the restrictions.

By addressing the crucial problem of separating real biometric samples from fake or artificial copies, fingerprint liveness detection makes sure that the system authenticates the submitted fingerprints. By stopping unauthorized access, this procedure is crucial for preserving the integrity of biometric systems. However, biometric cryptographic systems use biometric information like fingerprints to strengthen security by using cryptographic methods. This way, biometric identifiers are safely encrypted and shielded from unwanted access. The interdependence of liveness detection technologies and the efficacy of biometric cryptography systems highlights the relationship between these two domains. Robust liveness detection essentially acts as a cornerstone supporting the security of biometric cryptography systems, emphasizing the need for integrated solutions that handle biometric sample authenticity as well as biometric data management securely.

2.2.2. *Biometric Cryptosystem*

Biometric cryptosystems are mechanisms that enhance the security of biometric systems by embedding biometric templates with secret keys or auxiliary data. This approach ensures protection against malicious use and data tampering. Figure 4 shows the block diagram of biometric cryptosystem, which consist of two parts one is for enrolment where template are generated from the fingerprint images and stored in the dataset and the second part is authentication where template is generated from a query fingerprint and matched with the stored template to provide access to a system.

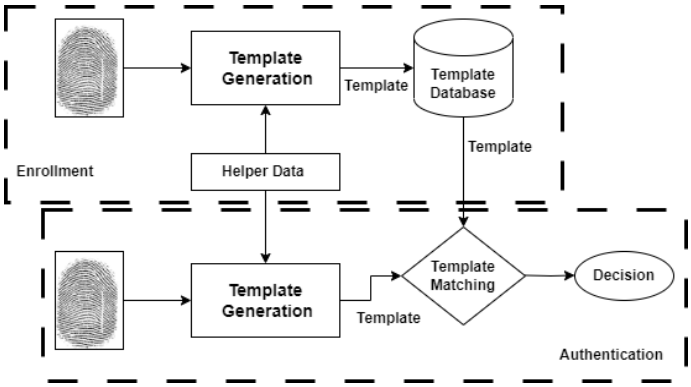


Figure 4: Biometric Cryptosystem

Pair-polar minutiae structures were used by [59] to present an alignment-free fingerprint cryptosystem. Their approach achieves robust security and improves privacy by changing minute structures. It has been tested on several databases and performs better than other systems in terms of Genuine Acceptance Rate (GAR) and False Acceptance Rate (FAR); nonetheless, a security risk arises from its dependence on the complexity of brute force attacks. In order to maximize security and efficiency, the authors in [60] suggest a feature-level sequential fusion algorithm for biometric cryptosystems. Experiments conducted on a finger vein database show that the algorithm performs better than the OR rule, decreasing input needs and obtaining a 1.47% False Acceptance Rate. Although the paper admits its limits in analysing some external security concerns, it shows that it is resilient to well-considered attacks. An ECC-free biometric key binding approach employing graph-based Hamming Embedding (GHE) and minutia-vicinity decomposition (MVD) for fingerprint minutiae-based representation was proposed by [61]. The strategy obtains GAR of 89% (FRR=11%, FAR=0.16%) for DB1 and GAR of 97% (FRR=3%, FAR=0.061%) for DB2, when tested on the FVC2002 datasets. The method balances security and performance trade-offs, but it is still susceptible to privacy attacks such as ARM and SKI, and it can only match fingerprint photos of the same finger taken under various circumstances or at different times. In [62], the authors provide a method to improve biometric template security that is based on a 2D logistic sine map (2DLSM). The technique employs diffusion and confusion processes, demonstrating effectiveness against a range of threats, by using chaotic streams from

2DLSM. Effectiveness has been evaluated on iris and fingerprint templates; nevertheless, real-time template production problems and noise sensitivity are among the constraints. A safe cryptographic authentication approach utilizing biometrics and the discrete logarithm problem was suggested by [63]. The approach produces cryptographic keys using a function F using iris codes, which allows it to handle mistake bits and achieve high accuracy rates. Promising results are shown by the CASIA iris database evaluation, which also notes that iris codes still require work in order to be improved in the future. BioKEY, a biometric-based cryptographic key generation technique based on convolution coding principles, was promoted by [64]. Unique cryptographic keys are created without saving templates by using fingerprint characteristics and applying convolution coding to specific locations. Tests conducted on common fingerprint databases such as MIAS, FVC2002, and FVC2004 show good accuracy and effectiveness; nevertheless, there are certain drawbacks such as difficulties with noisy or hazy fingerprints and a higher processing overhead in comparison to other techniques. ECC-based mutual authentication strategy for Smart Grid communications with biometric authentication integration for improved security and privacy was proposed by [65]. To safeguard against different types of security risks, the protocol consists of start-up, registration, and authentication stages. Notwithstanding its shortcomings, which include parameter sensitivity and vulnerability to certain assaults, the system shows efficiency through communication and computing cost comparisons with current protocols even if it does not offer precise accuracy measurements. In order to safeguard templates in multi-modal biometric systems, the authors in [66] provide a fuzzy vault technique that uses encryption to keep authentic templates safe from copycats. The approach satisfies optimal biometric security requirements with low FAR (0.1062) and zero FRR when tested on a simulated face and fingerprint database. In order to get the best security and accuracy, the authors recommend avoiding function creep attacks and balancing key length with chaff points. Hyper elliptic curve cryptography (HECC) for template encryption was used by [67] to introduce an improved iris recognition technique that achieves great security and accuracy. With the use of fuzzy logic matching and 2D Gabor filter for feature extraction, the system achieves 99.74% maximum accuracy and short identification time. Superior performance with reduced FAR, FRR, and EER measures is demonstrated by evaluation on the IITD and CASIA Iris V-4 iris datasets. A multi biometric cryptosystem for client-server network authentication is presented by [68], guaranteeing computation security and privacy. The approach enhances efficiency by achieving equivalent accuracy with less user input by using iris and fingerprint modalities from a dataset of 100 participants. The authors request protocol enhancements for increased security and verification speed, acknowledging the limitations in accuracy caused by biometric variability. A multi-biometric template security technique based on graph creation was proposed by [69] for cloud authentication. A branching factor graph with a low Equal Error Rate (ERR) of 0.66% is created using fingerprint and palm print features. The authors point out important benefits over conventional authentication techniques while also acknowledging limitations with regard to the encryption process and the quality of the input sample. Using the Diffie-Hellman algorithm and minute characteristics, the authors in [70] proposed a fingerprint-based crypto-biometric system for secure communication. Tested using FVC2002 and the NIST special database 4, the system reports metrics such as FAR, FRR, GAR, and EER, demonstrating privacy protection. The authors have acknowledged the restrictions pertaining to the irreversibility of biometric data, possible data distortions, and key creation. A multi-modal biometric cryptosystem utilizing fingerprint and ear characteristics was promoted by [71], which included preprocessing, feature extraction, and classification stages. The approach achieves excellent accuracy (98.76%) with a dataset that includes photos of ears and fingerprints. Performance is analysed using metrics such as sensitivity, specificity, accuracy, false positive rate, and false negative rate. The lack of multi-modal databases and the dependence on multi-modal systems because particular biometrics, such as ear characteristics, are unreliable are among the limitations. However, the method offers an efficient fix for identity and security systems. A biometric cryptosystem using random projection and back propagation neural network (BPNN) for template protection is introduced by [72]. Original biometric traits are transformed into unlinkable projected vectors by random projection, and BPNN is then used to map these vectors to secure keys. Better security and performance than current techniques are demonstrated via experimental assessment on a variety of biometric datasets. The authors stress the necessity for more investigation into novel ciphers for biometric template security and the improvement of security through multiple-biometric template protection. A novel fingerprint biometric cryptosystem using fuzzy commitment and CNN-based automated texture descriptor discretization was presented by [73]. Tests conducted on the FVC2000 DB2-A database show encouraging outcomes: 1.25% for FAR, 1.15% for FRR, and 2.83% for EER. The restricted number of photos per identification class presents challenges, such as precisely identifying reference points and producing a bigger training set. An effective cancelable biometric authentication framework using a Genetic Encryption Algorithm (GA) for increased security is presented by [74]. The process entails using GA for increased security, choosing the best sub-images, and permuting biometric templates. Promising findings were obtained via evaluation on a variety of datasets, such as the face and fingerprint datasets; nevertheless, more validation on bigger and more diverse datasets is required to ensure generalizability. A cancellable biometric security system using sophisticated chaotic maps to improve fingerprint identification was provided by [75]. Using various chaotic maps, chaos-based picture encryption produces convolution kernels that are then used to construct encrypted biometric templates. With an EER of 0.593% and a high detection probability of 96.139%), the augmented quadratic map 3 performs best in the system. The invertibility of biometric transformations and the requirement to investigate cancellable recognition in alternative biometrics are among the limitations. The Cancelable Biometrics Vault (CBV) was introduced by [76], who used winnowing and chaffing to create safe biometric templates for cryptographic key encoding. Experimental results demonstrate that, regardless of key size, the CBV's use of an extended BioEncoding scheme is acceptable for bit strings such as iris-codes, and that its

decoding accuracy is equivalent to that of the underlying CB construct. The reliance on an appropriate CB scheme for the biometric representation and computational complexity unfit for real-time applications are among the limitations. A multi-biometric cryptosystem employing Modulus Fuzzy Vault, augmenting input data with BDPHE and segmenting pictures with MBIRCH, was proposed by [77]. Features are extracted by the bidirectional deer hunting optimization algorithm, fused using score level and normalized feature fusion, and then safely stored. The approach enhances revocability and security while producing superior ROC, FAR, FRR, and GAR metrics; nonetheless, it has drawbacks related to algorithm focus and dataset authenticity. A machine vision gait-based biometric cryptosystem using a fuzzy commitment strategy was invented by [78]. Robust bits are chosen for the fuzzy commitment scheme after gait characteristics are retrieved using LTP and GEI. Tested on the CASIA A and CMU MoBo datasets, obtaining 0% FAR and FRR under certain scenarios. One limitation is that it might be challenging to identify people with complicated backgrounds or occlusions. An asymmetric cryptosystem integrating the elliptic curve method and optical scanning cryptography (OSC) was proposed by [79]. Using ECC from biometric pictures, the approach encodes things into holograms, guaranteeing excellent decryption accuracy and key sensitivity. The technique has to be optimized for speed and handles vulnerabilities to ciphertext-only assaults, even though security and key management have improved. In order to pre-process biometric images, the authors in [80] promoted the use of a fuzzy extractor that leverages deep learning, together with code-based cryptosystems to generate robust keys utilizing face biometrics. Promising accuracy metrics are found while evaluating the LFW and CelebA datasets. There is need for improvement in data privacy and security, as evidenced by limitations such as storage requirements and attack vulnerability. ElGamal encryption and Shamir's secret sharing are two cryptographic algorithms that are used in the blockchain-based user re-enrolment approach for biometric authentication systems proposed by [81]. Key security characteristics are satisfied by simulations that show secure re-enrolment in a matter of seconds; nevertheless, scalability and adversarial assumptions are acknowledged limits. More research is needed to determine the implementation specifics and practical viability. Based on Lagrange's interpolation, the authors in [82] provide a secure multi-biometric template protection method that uses irreversible and unlinkable image transformation. By utilizing feature-level fusion, the technique combines the properties of the fingerprint, iris, and palm print, attaining a high accuracy of 99.9816% when applied to high-resolution picture datasets. Issues include growing database sizes and possible departures from ISO/IEC 24745 standards, and privacy protection is still a problem. Biometric attributes can be randomly shuffled using a 3D chaotic map, as proposed by [83], to create a safe cancellable biometric cryptosystem. Larger datasets and other attacks are needed for more research, however evaluation on the ORL, FVC, and LFW datasets demonstrates encouraging results with very low Equal Error Rates ($6.2460x10^{1}3$) and high average Area under the ROC curve (0.9998). A robust cancellable biometric authentication technique is introduced by [84], which makes use of DNA sequencing theory, PWLCM, logistic map, and 3D chaotic maps. Through dispersion and confusion, the technique produces biometric patterns that are completely undefined while achieving increased security. Although computational complexity is still a barrier, evaluation on a variety of face and palm print datasets shows encouraging results in terms of AROC, FAR, DH, SSIM, and PSNR. An Enhanced Biometric Cryptosystem (BCS) using iris and ear modalities with Binary Robust Independent Elementary Feature (BRIEF) was introduced by [85]. Although vulnerable to database-level assaults, experimental assessment on AMI and UBIPr databases shows better performance across several parameters compared to state-of-the-art approaches, suggesting its applicability for security applications. Using deep learning and crypto-mapping, the authors in [86] provided a multi-biometric secure-storage technique that generated cancellable biometrics from fused pictures of the face, fingerprint, iris, and palm. While ICUB results reveal greater error rates, experimental results on the CASIA V4, MICHE, and MobiFace datasets demonstrate good AUCROC and low EER. The limitations mentioned by the authors include the necessity to strike a compromise between security and recognition performance, hardware requirements, and potential over fitting. In an effort to improve security, the authors in [87] proposed a multi-round AES cryptosystem with hierarchical hardware pipelined structures and biometric key generation. The system outperforms traditional AES systems in terms of space and energy efficiency because to resource sharing and simultaneous XOR operations. The undefined dataset size and the restriction to area and energy efficiency measurements are among the limitations.

Biometric characteristics are combined with cryptographic methods in biometric cryptography systems to protect authentication procedures, guarantee the privacy of biometric data, and prevent unwanted access. However, if the data is hacked, the intrinsic uniqueness of biometric features presents privacy problems. Cancellable biometrics are useful in this situation. In order to reduce the risk of prolonged exposure in the event of a data breach, cancellable biometric approaches convert biometric data into a non-reversible, pseudonymous form that can be updated or revoked as needed. The security architecture gains an extra layer of defence that improves user privacy and adaptability by incorporating cancellable biometrics into biometric cryptography systems. This addresses the reliability of authentication as well as the flexibility of maintaining biometric identifiers. This interaction emphasizes how crucial it is to combine cancellable biometric techniques with cryptographic security to offer a thorough and reliable solution to biometric data protection.

2.2.3. Cancellable Biometric

Cancellable biometric templates are introduced to enhance the security of biometric authentication systems by storing irreversible transformed versions of templates rather than original biometric templates. This approach prevents attacks and vulnerabilities while maintaining certain advantages, such as non-revocability. Techniques like mixing mechanisms and many-to-one transfer functions are commonly employed for generating cancellable biometrics. A basic block diagram of cancellable biometric is represented in Figure 5 where template is generated using transfer pro-

cess, if by any reason the template get compromise then it is very easy to replace the old template with a new one by changing the transformation process. Retaining identifiable biometric information despite transformation efforts aimed at unlinkability, potential mismatches due to lack of adaptation to temporal changes, increased computational demands, privacy risks from potential reconstruction, and usability issues regarding recognition reliability are some of the challenges associated with using high-quality initial biometric images for generating cancellable templates. In order to guarantee the security and usability of cancellable biometric systems in actual applications, it is imperative that these variables are balanced.

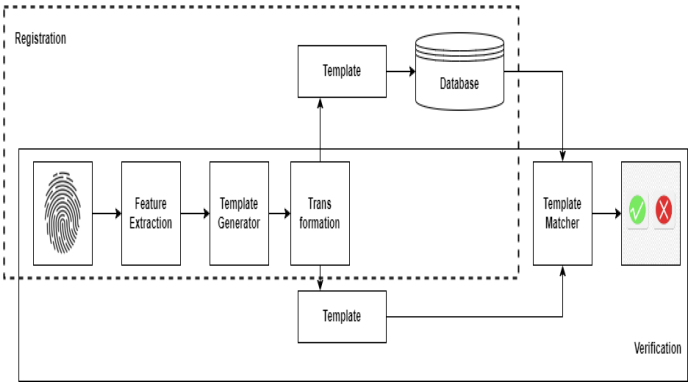


Figure 5: Cancellable Biometric

A fingerprint template protection technique utilizing fused feature-level structures that produce cancelable templates was presented by [88]. Improved performance was seen with EER ranging from 1.6% to 17.6% in the FVC 2002 and 2004 datasets. Two drawbacks are the expense of computing and the invertibility brought on by fusion. In order to take advantage of flaws in partial fingerprint-based authentication systems, the authors in [89] proposed creating "MasterPrints," which are artificial or actual partial fingerprints that match recorded templates for a large number of users. Their method proved that it was possible to impersonate users when tested on the FingerPass DB7 and FVC2002 DB1-A datasets; in some situations, Synthetic MasterPrints performed better than Sampled MasterPrints. Limitations include imbalanced datasets and limited application to minutiae-based systems, which have prompted more study into remedies. A one-factor cancellable biometric authentication technique using Indexing First Order hashing was suggested by [90]. It was assessed for accuracy performance, non-invertibility, renewability, and unlinkability. Tested on six datasets from the FVC 2002 and 2004 databases, yielding metrics for genuine-imposter distribution and Equal Error Rate (EER) that are adequate. Although it separates IDs from biometric templates, it assures unlinkability and is susceptible to standard symmetric key cryptosystem assaults. By using OIOM hash and MSH to create a pseudonymous identification, the authors in [91] promoted a one-factor cancellable palmprint recognition technique that achieved a recognition accuracy of 98.07%. although no specific limitations or restrictions are mentioned, experiments were conducted using the PolyU and TJU palmprint databases. A global multi-biometric system using deep neural networks for cancellable feature creation was proposed by [92], which achieved good performance on iris datasets (IITD Iris and MMU2). The technique acknowledges constraints in sensitivity and adaptability and blends dimensionality reduction, adaptive fusion, and revocability. Aspects of security and privacy are explored, and some adversarial threats are illustrated. In order to provide maximum safety of sensitive data, the authors in [93] provided the Secure Triplet Loss method for training end-to-end deep learning models to build non-invertible and unlinkable biometric templates. The approach is tested on facial recognition and ECG tasks, showing that it can successfully modify pre-trained models or create secure models from scratch. The dataset utilized for experimentation is not stated, despite constraints being addressed, such as sensitivity to certain assaults and the effect of demographic characteristics on system accuracy. Constrained Optimized Similarity-based Attack (CSA) was introduced by [94] as an improvement over earlier similarity-based assaults. It incorporates algorithm-specific limitations to optimize preimage production for impersonation. The usefulness of CSA is demonstrated through experiments using the labelled Faces in the Wild (LFW) dataset and Index-of-Max (IoM) hashing. The performance measures that are measured include the total success rate, the False Acceptance Impostor (FAI) rate, and the True Acceptance Impostor (TAI) rate. The success of CSA is contingent upon data leakage and constraint identifiability, which may restrict its application in certain biometric systems. In order to prevent unwanted access, the authors in [95] developed a non-invertible cancellable fingerprint template approach based on Delaunay triangulation of minutiae points. When evaluated against the FVC2002 database, the findings show promise in terms of identification accuracy, resilience against fingerprint distortion, and comparison to modern approaches. However, more research is necessary for wider validation due to concerns including hacked acquisition equipment and dataset size restrictions. A safe method for building fingerprint templates that are optimized was disclosed by [96]. This technique uses the quality of minutiae points to generate 3D shell-shaped templates. Tested on nine fingerprint datasets, the method yields a 0% Equal Error Rate (EER) and 100% accuracy in separating authentic individuals from imposters. Nevertheless, low-quality photos could be limiting its performance, indicating a possibility for improvement through multimodal biometric system integration. Using IoM hashing and BioHashing, the authors in [97] developed the first constrained-optimized similarity-based attack (CSA) against cancellable biometrics (CB). By optimizing preimage creation through algorithm-specific limitations, CSA outperforms Genetic Algorithm enabled similarity-based assaults (GASA). Success rate (SAR) and false acceptance index (FAI) of CSA, assessed on the LFW dataset, demonstrate the efficacy of CSA in breaking IoM hashing and BioHashing security. However, there are drawbacks to CSA due to its constant model complexity and dependence on hash code size. A cloud and Internet-of-things-ready method for cancellable biometric template generation was developed by [98]. By using the Greatest Common Divisor (GCD) between hazy biometric pictures, it guarantees authentication accuracy and non-recoverability. It is tested on several biometrics and achieves low EER values (down to 0.04%) and high AROC values (up to 99.59%), while it is recognized

to have limits with respect to pre-defined distortions and picture quality. A cancellable biometric authentication system with feature-adaptive random projection that includes feature extraction, transformation, and encrypted domain matching was presented by [99]. EER, GAR, and FAR metrics evaluation on the FVC2002 and FVC2004 databases revealed competitive performance. The need for more discriminating feature descriptors and investigation across a range of biometric features are acknowledged constraints. In order to provide biometric data safety and reliable authentication, the authors in [100] suggested a user authentication and key agreement system utilizing cancellable biometrics and PUF in multi-server settings. Their methodology beat state-of-the-art techniques when tested on the LFW dataset and assessed using the CMC, ROC, DIR, and DET curves. Recognized drawbacks include the requirement for assessment against complex assaults and scalability issues with larger data sets, indicating potential directions for further study. BioCanCrypto, a biocryptosystem using fingerprint cancellable templates that combines biocryptosystems and cancellable biometrics, was proposed by [101]. It takes cancellable templates and extracts cryptographic keys using a reusable fuzzy extractor with LDPC coding. tested using the FVC2002 dataset, showing encouraging outcomes despite low FRR and EER. The study is restricted to fingerprint data, though, thus it might be worthwhile to investigate alternative feature spaces or data modalities. A unique biometric template protection approach incorporating watermarking and cancellable transformation was suggested by [102]. The technique uses pair-polar coordinates with cancellable modification of minutiae and binary watermarking obtained from fingerprint minutiae. tested on the BioSecure and FVC2002 DB1 datasets, showing better EER and resilience to assaults than previous approaches. Investigating several biometric modalities to improve accuracy and robustness is one of the future research priorities. Absolute Value Equations Transform (AVET), a non-linear projection technique that ensures irreversibility by depending on the Absolute Value Equations issue, was promoted by [103]. AVET surpassed state-of-the-art approaches in bimodal circumstances and produced competitive performances in uni modal settings after being evaluated on eight datasets for different biometrics. A fixed sample size and susceptibility to brute-force attacks are two limitations. Using deep fusion and deep dream, the authors in [104] propose a multi-biometric cancellable system (MBCS) that creates tamper-proof templates using fingerprint, finger vein, and iris biometrics. It used a dataset with nine pictures per modality, outperforming comparable algorithms in EER, FAR, FRR, and AROC. All quantitative evaluations showed positive results. Larger datasets, high computing demands, and a drawn-out enrolment procedure are among the acknowledged limits. A cancellable multi-biometric identification system using ACM encryption and decimation to combine biometric data into a single template was presented by [105]. Performance measures, which were assessed using FVC 2002 and ICE 2005, included EER and ROC curve analysis. Its stated drawbacks were the need for better parameter estimate techniques and bigger databases. CSMoFN is a revolutionary cancellable multimodal biometrics system that combines face and periocular data, as reported by [106]. It uses pairwise angular loss and ArcFace for training, and it achieves an EER of 6.67% on average across six datasets and 2.12% on Facescrub. The authors warn against potential CB template inversion concerns while highlighting dual template-changeability and acknowledging difficulty in fair comparison. A new cancellable multi-biometric system combining deep learning-based fusion and selective encryption was developed by [107]. AES encryption, PRNG matrix XORing of a chosen ocular picture, and Viola-Jones facial segmentation are all used. Although constrained by the size of the dataset, it surpasses previous efforts with high entropy and low correlation, quicker enrolment, and better metrics like EER and AROC, indicating that medical imaging application is a promising area for future study. A unique cancelable biometric solution using deep learning for fingerprint and face biometrics on smart phones was presented by [108]. It improves security for Internet of Things applications by using Siamese networks with visual and text encoders. Prolonged testing shows improved performance compared to earlier approaches, surpassing ciphering-based systems in EER, FAR, FRR, and AROC, and attaining high accuracy metrics. MBFH, a unique cancellable biometric method that makes use of safe hashing and non-invertible transformations, was invented by [109]. Multi-biometric acquisition, feature extraction, Multi-Exposure Fusion (MEF), and SHA-3 hashing are all included in the process. Tests using retina, finger veins, palm, and dorsal vein pictures show high pairwise distances and good performance in generating text and visual templates, indicating flexibility for further improvements such as adding white Gaussian noise. A biometric template protection system for Euclidean and Cosine metrics was introduced by [110]. It used distance-preserving, one-way, and obfuscation modules with location-sensitive hash functions. Effectiveness against similarity-based and linear inequality attacks is demonstrated by evaluation on face datasets such as AR, CASIA, ORL, and LFW. However, this approach is not always the best option and requires pre-processing, which raises the computational costs significantly. Despite defences against assaults, the authors advise against assuming complete security in any situation. Using the chaotic Baker map to encrypt biometric templates, the authors in [111] provide a chaotic-based cancellable face recognition system that is highly accurate (98.43%) and adaptable to a variety of databases. The approach takes care of lighting, occlusion, and emotions, but it also recognizes the necessity for environmental adoption and trade-off between performance and user privacy. A cancellable biometric authentication system using image style transfer is proposed by [112]. During registration, users supply a face picture and a key image to create and store a template. In order to get high AUC values (>0.9) in most circumstances, authentication entails comparing freshly created templates with stored ones. Correlation coefficients and ROC curves are used to evaluate the results. One of the limitations is the possibility of authentication using marginally similar key images. This suggests that future research should put limitations on key image parameters.

Widely utilized in forensic and law enforcement settings, AFIS faces unique security issues such data breaches, spoofing attempts, and system flaws that could jeopardize the integrity and accuracy of fingerprint matching. In a similar vein, these vulnerabilities also affect general-purpose user authentication systems, which use fingerprints for iden-

tity verification and access control in common applications. Consumer-grade fingerprint authentication systems can benefit from and benefit from the protective tactics included in AFIS security measures. Developers can better address similar risks in general-purpose systems by knowing the vulnerabilities and countermeasures relevant to AFIS. This will help to ensure that strong security mechanisms are put in place to protect user data and prevent unwanted access. In order to improve overall security and user trust in biometric technology, it is crucial to transfer high-security solutions from specialized applications to more commonly used fingerprint authentication systems.

3. General Purpose User Authentication

Automated Fingerprint Identification Systems (AFIS) commonly rely on minutiae-based fingerprint matching due to its strong evidential value, versatility, storage efficiency, and reliable matching performance. However, an emerging alternative approach in AFIS is pore-based fingerprint matching, which offers distinct advantages. Pore matching techniques can be broadly categorized into two main methods: (i) pore alignment-based matching and (ii) direct pore comparison-based methods. These techniques harness the unique characteristics of pores within fingerprints to enhance the accuracy and robustness of authentication systems. Figure 6 represents fingerprint image enhancement and reconstruction, where first the image is normalized, then ridge orientation is detected. After the ridge orientation detection ridge reliability is estimated and the frequency of the ridges are calculated. The above two step helps to find the region of interest and then the region is passed through Gabor filtering to remove noises to construct a higher quality fingerprint image from a low quality fingerprint image.

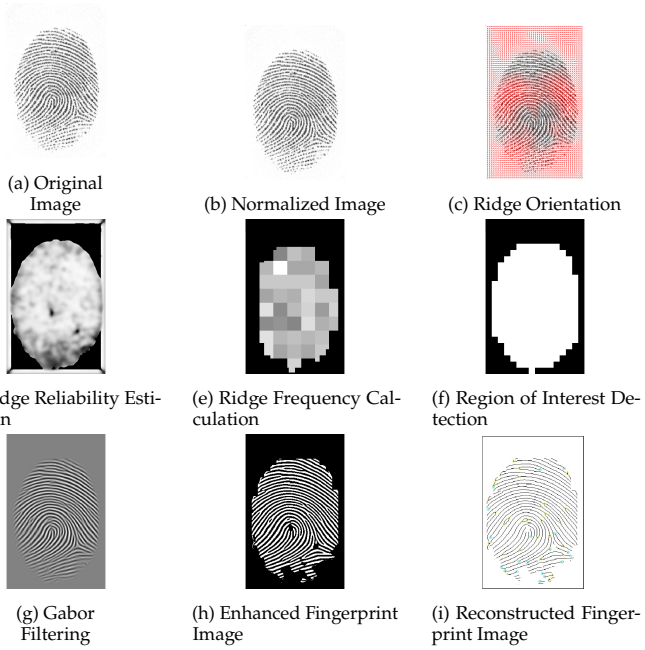


Figure 6: Fingerprint Image Enhancement & Reconstruction

Using global models, local analysis, and combination approaches, the authors in [113] provide techniques to improve fingerprint orientation extraction, obtaining an EER of 0.206% on the FVC2006 DB2 dataset. In addition to parameter tuning, they seek to increase recognition accuracy by addressing the shortcomings of benchmark datasets through pre- and post-processing phases. Using the AMFM fingerprint model and binary ridge pattern generation, the authors in [114] proposed a method to reconstruct whole fingerprint pictures from discrete points. Experimental assessment on FVC 2002 datasets shows better results than state-of-the-art methods with >86% of Successful Match Rates at FAR=0.01%. Reduction in performance for templates with fewer detail points and sensitivity to orientation field estimate accuracy are some of the limitations. Using image quality assessment (IQA) for liveness detection spanning fingerprint, iris, and face recognition, the authors in [115] provide a software-based technique for identifying phony biometric features. With the use of 25 generic IQA features, the technique performs competitively when compared to state-of-the-art methods, exhibiting minimal complexity appropriate for real-time applications and classification error rates below 3% and 21.4% Half Total Error Rate (HTER) on fingerprint datasets. One limitation is that in order to defeat obfuscation attempts, access to the entire picture and the necessary processing resources are required. A local model-based fingerprint classification technique including core block extraction, area division, and classification phases that is capable of managing noisy and incomplete data was suggested by [116]. Assessment using the FVC 2000, 2002, and 2004 datasets demonstrates higher accuracy (96.7% and 96.5%) in comparison to current techniques; low-quality fingerprints are purposefully tested to gauge performance in difficult scenarios. Using FVC2002 DB1 as an example, the authors in [117] improved accuracy over previous techniques by using a two-stage process for partial fingerprint enrolment and a multi-scale texture-based A-KAZE strategy for matching. In light of the algorithm's reliance on enhancement techniques and small dataset circumstances, potential future research areas include synthetic fingerprint creation and resilience in real-world applications. One open-source and three commercial-off-the-shelf (COTS) extractors are used by [118] to evaluate fingerprint minutiae extraction performance and resilience against picture degradations. Assessment utilizes several measures on a dataset consisting of 40,000 artificially produced and 3,458 public-domain fingerprint photos. The evaluation recognizes obstacles such as noise, differences in finger location, and environmental conditions that affect the accuracy of the system. By merging global and local distance metric techniques, the authors in [119] proposed the Global-Local Distance Metric (GLDM) framework for improving bio-cryptosystem accuracy utilizing signature-based biometric features. Average classification error rates of around 7% and 17%, respectively, are obtained via experimental assessment on the PUCPR and GPDS-300 datasets. Challenges include limited positive samples for training and design restrictions in the signature system. A CNN-based pore extraction method for fingerprint pictures was provided by [120], making it easier to extract Level 3 features. Though it relies on high-resolution photos and has difficulties identifying pores in occluded or blurred images, the system outperforms previous algorithms and achieves excellent accuracy across touch-based, touch less, and latent fingerprint datasets. Future research may concentrate on creating matching algorithms for diverse fingerprint photos using derived pore properties. A new

technique for high-resolution fingerprint pore comparison using rotational invariant edge descriptors and Root-SIFT descriptors was introduced by [2]. EERs and FMR1000s for the FVC2002 datasets are 1.86% and 0.12%, respectively, according to the evaluation. The authors support direct pore comparison methods in spite of the possible loss of actual correspondences, acknowledging the drawbacks of comparing pores only on the basis of geometric distance. In order to improve ridge structures and remove false minutiae, the authors in [121] provide a fingerprint enhancement and reconstruction method based on orientation and phase reconstruction. Despite limitations such as age, dirt, and medical issues impacting picture quality, validation on the FVC2002 and FVC2004 datasets shows promise for enhancing interoperability among minutia encoders and matchers. DeepPoreID was created by [122], which achieved higher performance, especially with tiny picture sizes, by using deep convolutional networks for descriptive pore characteristics in fingerprint matching. Significant gains in EER and FMR100 are shown by evaluation on high-resolution fingerprint databases; nonetheless, there are certain drawbacks, such as a need for a sufficient number of pores and difficulties with occlusion. It is underlined that high-resolution fingerprint databases are available as open-source resources for more thorough testing. A homomorphic encryption-based fingerprint authentication system that guarantees access control while protecting biometric template data was investigated by [123]. Tested on FVC2002 DB2, the system shows an EER of 9.23%. The computational time issues that were found encourage research into more effective homomorphic encryption techniques. For a thorough assessment, more testing with bigger datasets and varied scenarios is advised. The Shark Smell Optimization (SSO) technique was used by [124] to study a fingerprint authentication system. Achieved FAR of 0.00%, FRR of 0.00666%, and CVR of 99.334% after evaluation on a dataset of 150 student fingerprint pictures. The size of the dataset, lack of variety, and exclusive dependence on the SSO algorithm are limitations. Future research into more swarm algorithms and bigger datasets is advised to get better results. Using fully homomorphic encryption (FHE) and the TFHE library, the authors in [125] created an effective fingerprint authentication system that preserves privacy while guaranteeing safe processing and storage of fingerprint data. Tested on NIST Special Database 9, the system reaches fingerprint matching in an average of 166 seconds, acknowledging the time consumption of the bootstrapping procedure and suggesting improvements for the future. BioSec, a biometric authentication system using fingerprint authentication for secure communication among edge devices, was introduced by [126]. AES-128 bit encryption is used to protect biometric templates while they are being sent and stored in databases. Assessed using the FVC-2004 DB3 dataset with 70% accuracy, they recognize the limitations of symmetric encryption vulnerability and propose improvements to privacy mechanisms, study of lightweight encryption methods, and investigation of other biometric markers such as iris for enhanced security and usefulness. In [127], the authors invented a blind and reversible fingerprint image watermarking method using the differential method and DCT domains. Assessed on the FVC2002 fingerprint database, the technique employs DCT-transformed sub-vectors to incorporate watermark bits, facilitating direct watermark access throughout the extraction procedure. Evaluation measures included matching score and PSNR, which demonstrated a respectable level of fingerprint image security maintenance. The authors pointed out drawbacks like as noise sensitivity and compression, indicating that a bigger sample size would be necessary to increase performance. A set of two fingerprint matching methods, employing mtriplets and cylinder codes, fused with a supervised classifier for latent fingerprint recognition, was introduced by [128]. Achieved Rank-1 identification rates of 74.03% and 71.32% were evaluated on NIST SD27, GCDB, and MOLF DB4 databases. This work improves upon earlier approaches by evaluating matches using cumulative match characteristic (CMC) curves and plans to investigate other variables in further research for more accurate matching characterisation. Using attention mechanisms and Monte Carlo drop-out, the authors in [129] improved upon previous approaches by introducing an explainable fingerprint ROI segmentation model. High Jaccard similarity and Dice score measurements were obtained by testing on FVC datasets. Limitations include misclassification of background noise and performance changes caused by the size of the sample. An unsupervised technique for evaluating fingerprint quality based on minutiae detection confidence is called MiDeCon, and it was proposed by [130]. tested on FVC 2006 datasets; verification error and error-vs-reject curves showed superior performance compared to NFIQ1 and NFIQ2. The study did not address the limitations and limits of MiDeCon. ASRA, an Automatic Singular Value Decomposition-based Robust Fingerprint Image Alignment technique, was proposed by [131]. When tested on the FVC2002 and FVC2004 databases, ASRA demonstrated efficacy and efficiency in alignment, but with recognized shortcomings in terms of picture backgrounds and ROI extraction. For approach efficacy, more validation on bigger datasets is recommended. A unique fingerprint template protection and authentication method utilizing visual secret sharing (VSS) and super-resolution was suggested by [132]. The approach demonstrated security and resilience when tested on the FVC2002 DB1 dataset, while there are some acknowledged drawbacks, such as the requirement for higher-quality photos and the possibility of attacks on the super-resolution process. A dual-filter architecture and a novel texture descriptor, sDSIFT, were recommended by [133] to improve spoofing detection in fingerprint authentication systems. Tested using LivDet 2013 and 2015 datasets, the approach combines the dual-filter architecture with sDSIFT to achieve competitive accuracy. Nevertheless, shortcomings include the incapacity to distinguish between different filter kinds and spoofs created using materials that resemble real fingerprints. Three reliable methods utilizing fingerprint, iris, and voice characteristics for multi-modal biometric authentication were presented by [134]. The second technique, which used an SVM classifier and sum FFC, was evaluated on a dataset of 228 image and signal data, and it obtained a 100% classification rate; however, its scalability and real-world resilience were not evaluated. A secure online fingerprint authentication system with a cancellable fingerprint template design for privacy was provided by [135] for Industrial IoT devices over 5G networks. Tested on six public fingerprint databases, the system demonstrates competitive efficiency and performance, however it

recognizes the difficulties in preserving security while guaranteeing strong authentication over picture databases of poor quality. Using an Arduino Mega 2560 and a fingerprint sensor, the authors in [136] developed a fingerprint authentication system based on embedded systems. It generates templates using minutiae extraction and provides PIN and fingerprint combinations for smart home access. Although the accuracy is said to be high, low-end embedded systems' cost and durability against assaults present obstacles. CNN classification and Gabor filter-based feature extraction were used by [137] to investigate a fingerprint authentication technique. The approach beats other modern methods on the FVC2006 database, with a validation accuracy of 99.33%. Smaller training sets are still difficult to manage, despite results. A deep CNN model for fingerprint authentication was studied by [138], with 100% training and validation accuracy. One limitation is the diversity of datasets, which raises the possibility of strengthening robustness and generalizability. A multi-factor authentication approach was created by [139] to reduce False Acceptance Rates (FAR) and False Rejection Rates (FRR) in biometric systems. By combining several fingerprint sample permutations, their method increases system accuracy and fortifies it against impostor assaults. As opposed to single-factor authentication, experimental results on datasets like CASIA-FingerprintV5 and FVC2002 show lower mistake rates; nonetheless, the study admits several limitations, including addressing environmental deterioration and dataset size. InfinityGauntlet, a brute-force attack against smartphone fingerprint authentication (SFA) systems, was revealed by [140]. It uses SPI MITM to circumvent limit attempts and takes use of design flaws. They tested the assault on several devices using inexpensive equipment and a synthetic fingerprint generating technique, and they were able to detect four different types of attacks. One limitation is that security regulations demand the attack to be finished within 72 hours and that sensor hot plugging support be provided. A unified model for fingerprint authentication and spoof detection was developed by [141] using DualHeadMobileNet, a dual-head convolutional neural network. The model, using integrated datasets from FVC 2006 DB2A and LiveDet 2015, delivers low spoof detection error rates and good authentication accuracy. Computational effort and the requirement for bigger, more varied datasets for further studies are among the limitations. Blind-Touch, a privacy-preserving fingerprint authentication system using homomorphic encryption and machine learning, was introduced by [142]. Using PolyU and SOKOTO as benchmark datasets, the system attains great accuracy, with F1-scores of 93.6% and 98.2%, respectively. The influence of feature vector size on calculation time and the computational cost of constructing a traditional CNN with homomorphic encryption are among the challenges. The FingerPIN method, which combines PINs with fingerprints for multi-factor authentication, was proposed by [143]. According to a usability research, FingerPIN is more secure, quick, user-friendly, and efficient than conventional techniques. Restrictions include a limited sample size and a brief period of data collection, indicating the need for more study. An AVAO-enabled Deep Maxout Network (DMN) for fingerprint-based person authentication was proposed by [144]. The technique obtained an accuracy of 0.927 on the CASIA Fingerprint Image Database by employing the AVAO optimization algorithm of the method. The authors pointed out constraints such as the need for strong preprocessing and model training.

Fingerprints are being utilized more and more in general-purpose user identification to confirm identity on a variety of platforms, including secure facilities and mobile devices. This calls for quick and accurate verification procedures. Rapid fingerprint indexing and searching technologies are essential for fulfilling these expectations because they allow fingerprint records to be retrieved and matched quickly, improving system performance and user experience. The time needed for fingerprint verification can be greatly decreased by integrating sophisticated indexing and searching algorithms into general-purpose authentication systems, increasing the process's efficiency and scalability. This synergy emphasizes how crucial it is to optimize fingerprint search technologies to facilitate secure and seamless user authentication, guaranteeing that the speed and accuracy of the underlying verification procedures meet the convenience of biometric access.

4. Fast Fingerprint Indexing and Searching

In various application domains, the rapid response and high efficiency of an Automated Fingerprint Identification System (AFIS) are paramount. To achieve swift results from large datasets, a common approach involves using a combination of fast and accurate algorithms. This entails employing a fast algorithm to generate a subset of candidate fingerprints, followed by a more precise but slower algorithm for final matching. Efficient fingerprint indexing predominantly revolves around two strategies: (i) utilizing Level-I features like ridge orientation and frequency maps, and (ii) leveraging minutiae features. The primary aim of fingerprint indexing is to reduce search space and time, especially when dealing with massive gallery image datasets. This section highlights several algorithms that exemplify this approach. Figure 7 shows the fingerprint processing and matching in distributed environment where the features from the fingerprint image is extracted at the client side and passed it to a server, where multiple feature matchers run in parallel to match the query feature in large fingerprint database. After matching the result is again send back to the client side.

To analyse huge databases efficiently, the authors in [145] created a distributed fingerprint matching system. The system achieves flexibility and scalability by using a two-level distributed design. Studies using NIST and synthetic fingerprint datasets showed that, when processing nodes and thread counts rose, execution times improved while accuracy remained similar to that of conventional AFIS. By parallelizing database searches, the authors in [146] introduced a GPU-based fingerprint recognition system that achieves fast processing rates. Although the approach can quickly identify millions of fingerprints per second, its practical use is limited by its 2% mistake rate. While highlighting the promise of GPU-based algorithms, the paper also notes that more development and integration into hybrid systems are required to achieve higher accuracy.

Using locality sensitive hashing (LSH) and minute cylinder-code SDK for effective indexing with pose lim-

itations, the authors in [147] developed an absolute registration method for fingerprint indexing. Experiments using databases such as the NIST Special Database 14 and a law enforcement fingerprint database shown increased efficiency and accuracy. Among the drawbacks are the high dimensionality of minutiae descriptors and the requirement for additional study to improve resilience against noise and unpredictability. In [148], the authors presented a fingerprint indexing technique that uses convex core points and minutiae pairs, combining candidate list reduction and k-means clustering to achieve effective indexing. The methodology outperforms state-of-the-art methods on six datasets by using coaxial Gaussian track coding and MBP representation for precise matching. One limitation is that the hardware and software utilized in the trials varied, and there is a lack of information on the values of the parameters.

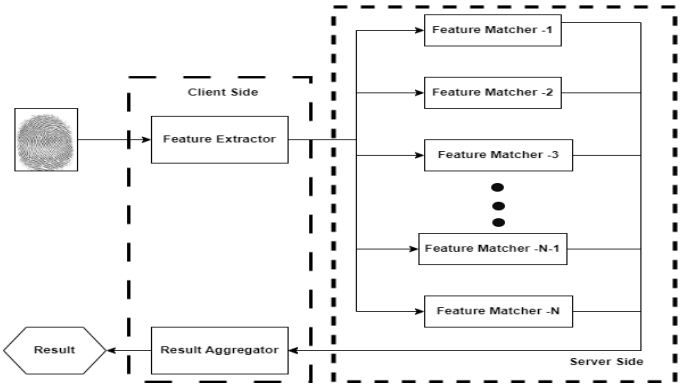


Figure 7: Fingerprint Processing & Matching In The Distributed Model

Improved retrieval efficiency was achieved by [149] using deep convolutional neural networks (DCNN) to build a minutiae-centred fingerprint indexing approach. They achieved superior results on five benchmark databases with their unique aggregating methodology that made use of 1-D CNN. Metrics for evaluation included penetration and error rates; higher penetration rates and lower error rates were noted. Two of the limitations were the high processing costs for MCC-based approaches and the low quality photos in FVC2000 DB3a. To address the shortcomings of minutiae-based approaches, the authors in [150] studied DeepPrint, a deep network for obtaining fixed-length fingerprint representations. Using the NIST SD4 and SD14 datasets, DeepPrint outperformed leading commercial off-the-shelf (COTS) SDKs in terms of accuracy and search speed. DeepPrint showed encouraging results in fingerprint identification, despite limits in benchmark saturation. CNNAI, a CNN-based technique for fingerprint recognition utilizing geometrical minutiae arrangements, was studied by [151], and it outperformed state-of-the-art algorithms in terms of identification rates. CNNAI demonstrated its abilities by achieving Rank-1 identification rates of 84.5% and 80% on the FVC2004 and NIST SD27 latent fingerprint datasets, respectively. Reliance on a minimum of eight minutiae points is one of the limitations, and using MINU-EXTRACTNET to extract all minutiae presents difficulties that might potentially impact identification rates. By developing a pore-based indexing technique for high-resolution fingerprints, the authors in [152] significantly reduced preselection error rates compared to previous approaches on a variety of databases. The approach outperformed state-of-the-art minutiae-based algorithms on DBI and IITI-HRFP, according to tests conducted on datasets comprising DBI, DBII, IITI-HRFP, and IITI-HRF. Restrictions include decreased efficacy on databases with high-resolution whole fingerprints, which motivates more study into distortion correction methods. A latent fingerprint indexing technique integrating minutiae-based features, global and local matching, and machine learning-based segmentation was suggested by [153]. Up to 94.77% accuracy was shown in experiments on the IIIT-D dataset, with indexing times varying from 5.29 to 25.92 seconds. Managing incomplete perceptions, background noise, inadequate ridge clarity, and nonlinear distortions are among the difficulties. An MCC-based indexing technique for latent fingerprint recognition was described by [154], beating competitors on the NIST SD4 and NIST SD14 databases by at least 1% and 3%, respectively. Although effective, it can have performance issues with different kinds of databases or with less-than-ideal detail extraction, and its temporal complexity prevents real-time applications. Employing graph-based algorithms for indexing and refining, the authors in [155] provided a pore-based fingerprint retrieval technique that outperformed previous approaches on databases DBI and DBII in terms of speed and accuracy. Its shortcomings, however, include the dependence of indexing precision on cluster accuracy and the computational cost that increases with the number of pores. Two fingerprint search strategies utilizing classic inverted index methods were proposed by [156]. They were examined on FVC2002 DB1a and a private dataset, with differences in document generation methodologies and minutiae handling. The second approach's scalability and sensitivity to low-quality minutiae were found to be limits, despite the fact that it produced reduced error rates. Using highly discriminative embeddings for constant-time identification and k-means or LSH index table generation, the authors in [157] present PalmHashNet, a palm print database indexing technique. Evaluation on four benchmark databases shows above 99% accuracy; nonetheless, there are several constraints, such as assumptions about the positioning of the palm area and unproven scalability on big datasets. By combining GIST descriptors from face, iris, and palm print biometrics, the authors in [158] developed an efficient multidimensional spectral hashing (MDSH) technique for data retrieval from multi biometric databases. Despite computational and memory constraints, experiments conducted on databases from IIT Delhi reveal enhanced accuracy measures that surpass tree-based indexing approaches and random hashing techniques. More efficiency improvements are proposed, such as lower memory use and processing costs, and investigating alternatives to local features. To lessen the computational burden in biometric identification systems, the authors in [159] proposed a nearest quality score-based intelligent search, which resulted in workload reductions of up to 38% for face, 31% for iris, and 29% for fingerprint databases. The technique, tested on the FERET, CASIA, and FVC2002 datasets, effectively indexes queries using sample quality scores that may be applied to a wide range of biometric features. Intraclass variability and susceptibility to various quality estimator techniques affecting the number of comparisons are challenges. FKPIndexNet is an innovative approach for finger-knuckle-print (FKP) identification, de-

veloped by [160]. It uses customized autoencoder networks and similarity-preserving hash codes. The PolyU-FKP and IITD FKP databases show 100% hit rate in experimental findings at low penetration rates, indicating that index table creation approaches have limitations. A signal-based fusion approach for indexing biometric databases was presented by [161], which reduced computational effort by up to 70% without sacrificing biometric performance. Promising results are obtained when evaluated on benchmark databases such as FERET and FRGCv2, however these need the recomputation of indexes for new enrolments and high-quality face photos. By utilizing cancellable methods and DNN-based embedding extractors, the authors in [162] developed a novel approach to multi-biometric indexing that preserves privacy while improving biometric performance by 57%. Evaluation using a composite dataset reveals advancements but also points up extractor compatibility and dataset representation constraints. In their investigation of AFR-Net, an attention-driven fingerprint recognition network, the authors in [163] combined CNN-based and attention-based embeddings for improved performance. By utilizing a variety of datasets, AFR-Net surpasses baseline models and attains excellent accuracy metrics in verification and identification tasks. The lack of domain expertise in fingerprints and the requirement for better latent fingerprint preparation methods are among the limitations. An effective Gravitational Search Decision Forest (GSDF) method for fingerprint identification was created by [164] by merging the gravitational search algorithm (GSA) and random forest (RF). Comparing the GSDF technique to conventional machine learning classifiers, it yields higher identification rates for both entire and latent fingerprints. The authors agree that although the technique beats current methods, additional tuning is necessary for the recognition of low-quality latent fingerprints.

Infant fingerprint identification presents a unique set of challenges compared to adult fingerprint identification, which has a well-established sector with reliable techniques and tools for collecting and evaluating mature fingerprint patterns. Because of their fingers' quick growth and development, infants' fingerprints are more prone to distortion and have less defined ridges, therefore effective identification requires specific methods and modifications. It is necessary to modify fingerprint capture and processing technologies that have been developed for adults in order to address these issues with newborns. Researchers and practitioners can improve the overall efficacy and accuracy of fingerprint identification systems across all age groups by linking the methodology and technology utilized in both fields. This will guarantee that the systems remain accurate and dependable from infancy through adulthood. For comprehensive fingerprint identification systems that meet the needs of a wide range of demographics, this integration is essential.

5. Infants' Fingerprint Identification

Emerging as a challenging application, infants' fingerprint identification plays a crucial role in maintaining accurate records of a child's vaccination history, nutritional needs, and overall identity throughout their lifetime. Among various biometric options, fingerprinting emerges as one of the most suitable choices due to its stability and uniqueness. While alternative biometrics like iris scanning or facial recognition are less feasible for infants due to practical challenges, fingerprints offer a reliable bio-marker. In the field of infants' fingerprint identification, limited research exists, but several notable contributions have been gathered from the literature. Figure 8 shows fingerprints of an infant at different age, which demonstrate the different density of ridges during different age of a infant.

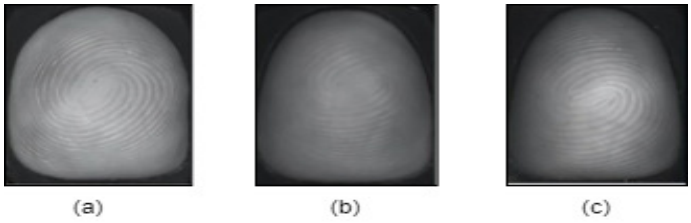


Figure 8: Fingerprints of an Infant at different age: a) 4 months b) 6 months c) 1 yr 3 months

A cost-effective, high-resolution fingerprint scanner for newborns was created by [165]. They achieved a rank-1 accuracy of 73.98% by building a database with pictures from 16,384 infants. In [166], the authors assessed the accuracy of child fingerprint recognition using NFIQ 2.0 measures in a follow-up research. Using a 500 PPI dataset, they were able to achieve TAR of 99.5% with FAR of 0.1% for children aged 12 months. An inexpensive newborn fingerprint identification system called InfantPrints was created by [167]. It consists of a high-resolution matcher and a specially designed fingerprint scanner. Assessed on a 315 baby longitudinal database, the system demonstrated excellent accuracy from enrolment at 2 months to authentication at 1 year. Among the drawbacks is the lack of a reliable automatic alignment technique for baby fingerprints. Using wavelet feature extraction and K-NN classification, the authors in [168] developed a baby's footprint identification system. Using a 200x500 pixel ROI and level 4 wavelet decomposition, the system obtained 99.30% accuracy with a dataset consisting of 600 footprint photos and 30 newborns. The modest sample size and the requirement for more testing on a bigger dataset for real-world application are acknowledged constraints. A newborn fingerprint identification technique using deep learning for fingerprint categorization was described by [169]. The approach obtained 78.4% classification accuracy compared to hand classification with a dataset consisting of 1,357 training photos, 166 validation images, and 1,181 test images. The requirement for significant computer resources and the difficulties in obtaining clear fingerprint pictures from some newborns are among the limitations. Using fingerprint, iris, and outer ear shape modalities, the authors in [170] created and assessed biometric recognition systems for babies. For the fingerprint and iris modalities, new hardware and software were created, but the outer ear shape modality made use of already-existing technologies. At a public clinic, data on newborns and young children under the age of one was gathered. FTA (failure to acquire) and equal error rate (EER) were used as accuracy measures. With an EER of 0.16%, the ear modality performed the best; nevertheless, inadequate reporting, inconsistent devices, and data

collection from newborns and young toddlers presented difficulties. A multi-instance contingent fusion approach for newborn fingerprint verification was presented by [171], which combined the left thumb and right index finger prints of infants ranging in age from one day to six months. The approach achieved 73.8%, 69.05%, and 57.14% accuracy for 1-month, 3-month, and 6-month intervals between enrolment and query prints, respectively, by using a 500 ppi fingerprint scanner and NIST feature extractor MINDTCT. Among the limitations were the requirement for fusion with increasing time intervals between photos and the lack of supplemental information in permission forms. M²BRTPC, a new Multi-modal Biometric Recognition for Toddlers and Preschool Children, is presented by [172]. Its goal is to identify children using minimum biometric characteristics such as fingerprints, faces, and iris. Their research, which acknowledged age range restrictions and the inherent difficulties in biometric identification, improved performance by $14.75e^{1}8.75\%$ using iris and fingerprint data from over 100 children ages 18 months to 4 years. Using a pre-trained ResNet-50 model, the authors in [173] investigated a fingerprint identification method for infants and toddlers, solving issues with image clarity and quality. The process entails cropping and enhancing images in order to precisely extract fingerprint traits. Using a dataset of 154 participants, the experimentation yielded an accuracy of 82.47%, a false rejection rate of less than 18%, and an authentication duration of about 2 seconds per fingerprint, while taking into account restrictions such as tiny fingerprint size and picture quality limits. Employing the CLCF dataset supplemented with a hybrid technique, the authors in [174] presented Child-CLEF, a CNN-based children detection system employing contactless fingerprints. The suggested Child-CLEF Net model outperforms current systems with 98.46% accuracy and 1.99% equal error rate by extracting minutiae and using BOZOROTH3 for identification. Larger datasets are required, as is the exploration of other enhancement and feature extraction techniques. Future developments should take patch-based newborn identification into consideration. These are some of the limitations.

Biometric identification systems for newborns and toddlers have advanced significantly in terms of accuracy and technological adaptability, but they still face significant obstacles in terms of data quality, moral issues, and practical implementation. Improving algorithms, expanding datasets, strengthening privacy protocols, and improving usability will be crucial in tackling these issues and boosting the dependability and relevance of these systems in paediatric healthcare and other fields.

A key component of biometric authentication is fingerprint identification, which uses distinctive ridge and valley patterns to confirm each person's identity. Though useful, it can have drawbacks such spoofing, ambient conditions, and differences in fingerprint quality. In order to improve accuracy and reliability, multi-modal identification systems integrate various biometric modalities, such as voice prints, iris patterns, fingerprints, and facial recognition. Multi-modal systems are able to complement fingerprint data with other biometric identifiers, making up for any one modality's shortcomings and offering a more robust and all-encompassing authentication procedure. This integration makes biometric systems more flexible and user-friendly while simultaneously enhancing overall security. Knowing how techniques might improve fingerprint recognition shows the possibility of developing more flexible and safe biometric systems that take advantage of several biometric properties.

6. Multi-modal Fingerprint Biometric

In [175], the author reported a multi-modal biometric system that combines face, fingerprint, and signature modalities with feature extraction techniques including Principal Component Analysis and Stationary Wavelet Transform. In order to address problems like spoof attacks and noisy data, they evaluate system accuracy on the YALE-FVC2002KVKR database by applying score and decision level fusion. Through the use of multi-modal strategies to overcome restrictions, the project aims to boost the dependability of biometric systems. In [176], the authors proposed improved uni modal and multi-modal biometric recognition systems using fingerprints and ECG signals using both traditional approaches and deep learning. When tested on virtual datasets like the FVC2004 fingerprint database and the MIT-BIH ECG database, multi-modal systems perform better than uni modal ones. Some of the drawbacks are the scarcity of ECG databases, the challenge of finding real-world datasets, and the high processing costs related to deep learning models. In [177], the authors have presented a deep learning-based multi-modal biometric fusion model that blends score, feature, and pixel layers to increase identification accuracy. Evaluation on a simulated dataset combining iris, fingerprint, and face data shows 99.6% accuracy using Euclidean distance metric learning and modality-specific network training for practicality. Limitations originating from the dataset's dependability may affect the accuracy rates. A 98.5% accurate HGSSA-bi LSTM model that combines fingerprint and iris biometrics is presented by [178]. When tested on the CASIA dataset, the model demonstrates excellent sensitivity and precision, but it also acknowledges the cost and complexity of multi-modal systems. In [179], the authors offers "Secure Sense," a multi-modal biometric system that incorporates face, fingerprint, and iris data and achieves 93% accuracy. The limitations of uni modal systems are addressed by decision-level fusion approach, which enhances strong authentication by utilizing real-time and web-based datasets.

7. Comparative Analysis

The Table 1 provides a comprehensive overview of unformatted fingerprint image authentication methods, encompassing various techniques aimed at enhancing the quality and recognition accuracy of latent fingerprints. Each method offers unique advantages and faces specific limitations, highlighting the need for ongoing research to address existing challenges and explore new avenues for improvement. Future research in this field should focus on several key directions. Firstly, there is a pressing need to develop robust and scalable algorithms capable of handling diverse and challenging latent fingerprints, including those with low quality, partial information, or distortion. This may involve further exploration of deep learning architectures, such

as convolutional neural networks (CNNs) and generative adversarial networks (GANs), to effectively enhance latent fingerprint images while preserving crucial details and minimizing noise. Secondly, efforts should be directed towards standardizing evaluation protocols and benchmark datasets to facilitate fair comparisons between different methods and promote reproducibility across studies. This includes the creation of large-scale, publicly available databases encompassing a wide range of latent fingerprint images captured under various conditions, which can serve as a common test bed for evaluating the performance of different algorithms. Additionally, research should address ethical considerations surrounding the deployment of unformatted fingerprint image authentication systems, including privacy concerns, algorithmic bias, and potential misuse of biometric data. Developing transparent and accountable frameworks for data collection, storage, and usage is essential to ensure the responsible and ethical implementation of these technologies in real-world applications. Furthermore, the integration of multi-modal biometric fusion techniques, such as combining fingerprint with other biometric modalities or contextual information, holds promise for enhancing the robustness and reliability of authentication systems, especially in challenging scenarios or under adversarial conditions. Lastly, advancements in hardware technologies, such as high-resolution sensors and efficient processing units, can significantly contribute to improving the accuracy and efficiency of unformatted fingerprint image authentication methods. Collaborative efforts between academia, industry, and regulatory bodies are essential to drive innovation, address emerging challenges, and ensure the continued advancement and responsible deployment of fingerprint authentication technologies in various domains. By pursuing these research directives, the field of unformatted fingerprint image authentication can continue to evolve and meet the growing demand for secure and reliable biometric authentication solutions.

Table 1: Comparative Analysis of Unformatted Fingerprint Image Authentication Methods

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[4]	2012	Robust alignment algorithm, descriptor-based Hough transform	NIST Special Database 27 (NIST SD27)	Rank-1 accuracy	53.5%	Performance highly correlated with minutiae count and print quality. Not effective for singular point-based alignment.
2	[5]	2014	Feedback mechanism from exemplar print	NIST SD27, WVU latent databases	Improvement in identification accuracy	0.5-3.5%	Trade-off between introducing feedback and system complexity. Effectiveness depends on exemplar print quality.
3	[6]	2016	Deformable Minutiae Clustering	NIST SD27, FVC2002, FVC2004, FVC2006, NIST SD4	Accuracy	Up to 85.6% (Cylinder-Codes), 83.3% (m-triplets)	Slow speed, scalability issues, not effective for latent-to-latent fingerprint identification.
4	[7]	2017	Adaptive latent fingerprint segmentation	NIST SD4, NIST SD 27, IITD CLF,	Rank 50 identification	78.7%,	Assumption of consistent ground truth across examiners, limitations of SIVV based metric.
5	[8]	2018	Collaborative filtering model for enhancing fingerprint image	FVC2004	EER, FMR100	4.54, 7.5	Fixed patch size limitation, sensitivity to input quality.
6	[9]	2018	Convolutional neural network (CNN) FingerNet for latent fingerprint enhancement	NIST SD27	Matching accuracy	47.7%	Small dataset size, lack of ground truth for region of interest.
7	[10]	2018	Automated latent fingerprint recognition system with ConvNets	NIST SD27, WVU latent databases	Rank 1 identification accuracies	64.7%, 75.3%	Poor ridge quality, background noise, dependence on manual ROI selection, long processing times.
8	[11]	2019	Minutiae-based matcher improvement using rare minutiae	GCDB	Rank-1 identification	92.72%	Need for manual intervention, dataset size limitations.
9	[12]	2019	End-to-end latent fingerprint identification system	NIST SD27, MSP, WVU, N2N, background set of 100K rolled prints	Rank-1 retrieval rates	65.7%, 69.4%, 65.5%, 7.6%	Challenges in cropping algorithm, marking minutiae, and scalability.
10	[13]	2020	Asynchronous processing for Latent Fingerprint Identification (ALFI)	NIST SD27, FVC 2004, FVC 2006	F1-score, Equal Error Rate (EER)	4.18%, 3.36%	Limited test set, single classification approach, not tested on non-fingerprint latent features.
11	[14]	2020	LQMetric: objective, automated tool for measuring the quality of latent fingerprints	NIST ELFT-EFS-2 evaluation	Clarity prediction, AFIS performance	61.4%	Performance not generally applicable to other AFIS algorithms or systems.
12	[15]	2020	Progressive GAN-based method for latent fingerprint enhancement	NIST SD27	CMC curve metrics	76%	Computationally expensive, limited dataset size, may not work well with very weak features.
13	[16]	2020	End-to-end automated latent fingerprint identification system using DCNN-FFT enhancement	FVC2002, FVC2004, NIST SD27	Precision, recall, F1 scores	100% for FVC2002 and FVC2004, 84.5% for NIST SD27	Computational time for minutiae extraction, limited evaluation dataset.
14	[17]	2020	Deep nested UNets architecture for automatic segmentation and enhancement	NIST SD27, IIITD-MOLF	PA, MPA, MIoU	0.96,0.96,0.84	Lack of publicly available databases with pairs of low-quality latent and high-quality fingerprint images.
15	[18]	2020	Non-minutia latent fingerprint registration method using dense fingerprint patch alignment	NIST27, MOLF	Deviation between mated minutiae, registration performance	87.22%, 2.42	Relying only on 2D information, computationally demanding.

16	[19]	2021	Fusion of pores and minutiae for latent fingerprint identification	IIITD Latent database	True detection rate, false detection rate	82.89%,21.2%	Limited database size, computational expense.
17	[20]	2021	Latent fingerprint identification using Ratio of Minutiae Triangles	FVC2004, NIST SD27	Rank-1 recognition accuracy	78.75%,86.82%	Handling of partial fingerprint, computational complexity.
18	[21]	2021	Hybrid model using EDTV for enhancement and Chan-Vese for segmentation	NIST SD27, WVU DB	Rank-1 identification	72%	Limited database size, effectiveness in handling complex latent fingerprints.
19	[22]	2021	Adaptive latent fingerprint segmentation and matching using Chan-Vese based on EDTV	NIST SD27	RMSE	0.1837301	Inadequate accuracy of existing techniques for segmentation.
20	[23]	2022	Automatic latent fingerprint identification system using scale and rotation invariant minutiae features	FVC2004, NIST SD27	Rank-1 identification accuracy	97.5%,88.8%	Handling partial fingerprints, computational complexity.
21	[24]	2022	Minutia patch embedding network (MinNet) model	EGM Test Dataset, FVC-Latent Test Dataset, Tsinghua Distorted	Rank-1 Acc	92.39%, 85.88%	Limited success with partial or severely distorted fingerprints
22	[25]	2022	Multi-scale fixed-length representation approach	Hisign, NIST SD27, MOLF, N2N	Rank-1 Accuracy	98.8%,99.81%	Performance may vary on simulated fingerprints, optimization limited to study scope
23	[26]	2022	Residual encoder-decoder architecture	IIIT-Delhi Multi Sensor Latent Fingerprint (MOLF) database	Rank-25, Rank-50 Acc	48.95%, 70.89%	Ridge-based, limitations in reconstructing some parts of images, evaluation on different databases needed
24	[27]	2022	Analysis of fairness in latent fingerprint prediction	FBI WVU BioCop 2008 database	AUC, SN, SP, PPV	65.7%,69.4%	Limited to FBI BioCoP database, quality measurements may not fully represent unbiased quality score, modelling assumptions of LFIQ algorithm may not always be satisfied
25	[28]	2023	Universal Latent Fingerprint Enhancer (ULPrint)	MSP database, NIST 302 database, Synthetic latent fingerprint	Rank-1 retrival	29.07%	Scarcity of latent ,Challenges like occlusion, background variation, some failure cases mentioned
26	[29]	2023	FingerGAN for latent fingerprint enhancement	NIST SD27, IIIT-Delhi MOLF	Rank-1 accuracy	76.36%	Computational complexity, absence of true mates
27	[30]	2023	Combination of local and global features with automatic seg.	NIST SD 27, NIST SD 302, MSP, MOLF DB1/DB4, MOLF DB2/DB4	Rank-1 Retrieval Rate	84.11%,70.43%,62.68%	Challenges like low contrast, occlusion, varying backgrounds, some failure cases mentioned
28	[31]	2023	Hybrid technique called ACSACO for latent fingerprint recognition	NIST SD-27	Precision, Recall, F-score	82.07%, 98.86%,89.68%	May not work well for complex backgrounds or overlapping latent fingerprints, need to explore other optimization techniques and datasets for validation
29	[32]	2023	Generation of synthetic latent fingerprints for data augmentation	NIST SD27 ,MSP latent Database	True Detection Rate	75.19%,77.02%	Use of only one pre-trained fingerprint matcher, lack of publicly available operational latent fingerprint databases

The Table 2 provides a detailed comparative analysis of various fingerprint liveness detection methods, encompassing a range of techniques such as feature extraction, machine learning, deep learning, and fusion approaches. Each method demonstrates different strengths and limitations in terms of accuracy, dataset applicability, and susceptibility to spoofing attacks. For instance, deep learning-based approaches, such as convolutional neural networks (CNNs) and stacked auto encoders, have shown promising results in achieving high detection accuracy. However, they often require large and diverse datasets for training to generalize well across different sensors and spoofing materials. Future research in fingerprint liveness detection should address several key challenges and explore new avenues for improvement. Firstly, there is a need to enhance the robustness and generalization capabilities of existing methods by leveraging larger and more diverse datasets that encompass various sensor types, image qualities, and spoofing materials. Additionally, researchers should focus on developing techniques capable of detecting low-resolution fingerprints and effectively handling imbalanced datasets to mitigate bias and improve overall performance. Secondly, the development of adaptive and resilient liveness detection systems that can dynamically adjust to evolving spoofing attacks is crucial. This may involve exploring novel approaches such as adversarial training, ensemble learning, and anomaly detection to detect and adapt to emerging threats effectively. Moreover, incorporating multi-modal biometric fusion techniques, such as combining fingerprint with iris or face modalities, could further enhance the robustness and reliability of liveness detection systems. Furthermore, research efforts should concentrate on addressing ethical considerations related to the deployment of fingerprint liveness detection systems, including privacy concerns, algorithmic fairness, and transparency in decision-making processes. Developing standards and guidelines for evaluating and benchmarking liveness detection methods could also facilitate fair comparisons and promote reproducibility across different studies. Lastly, advancements in hardware technologies, such as improved sensor designs and embedded processing capabilities, can play a vital role in enhancing the efficiency and real-world applicability of fingerprint liveness detection systems. Collaborative efforts between academia, industry, and regulatory bodies are essential to drive innovation and ensure the responsible development and deployment of liveness detection technologies in various domains, including cyber security, law enforcement, and mobile authentication. By addressing these research directives, the field of fingerprint liveness detection can continue to evolve and meet the growing demand for secure and reliable biometric authentication solutions.

Table 2: Comparative Analysis of Fingerprint Liveness Detection Methods

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[33]	2016	Combination of low-level features, shape analysis, and PCA	LivDet 2011, LivDet 2013	EER	3.95%	Need for more testing on different sensors and spoofing materials.
2	[34]	2017	Co-occurrence array-based feature extraction and SVM classification	LivDet09DB, LivDet11DB	Average classification error	6.2%	Loss of image information through quantization, large feature dimension.
3	[35]	2018	Deep CNN using local patches centred on minutiae	LivDet 2011, LivDet 2013, LivDet 2015	Average accuracy	99.03% (LivDet 2015)	Need for more diverse datasets and ethical considerations.
4	[36]	2019	BP neural network using difference co-occurrence matrices	LivDet 2013	Classification accuracy	5.65%	Lack of diagonal direction difference co-occurrence matrices, challenge in distinguishing poor quality images.
5	[37]	2020	Semi-supervised stacked auto encoder-based hierarchical feature learning	LivDet 2011, LivDet 2013	Average classification error	19.62%	Insufficient size of the dataset, difficulty in distinguishing poor quality images.
6	[38]	2020	Adversarial attacks on deep learning-based liveness detection models	LivDet 2013, LivDet 2015	Error rate, FAR, FRR	4.3%, 3.7%	Limited diversity of fingerprint databases, simplicity of fingerprint liveness detection models.
7	[39]	2020	Score-level fusion of fingerprint matching and liveness detection	LivDet2015, LivDet2019	Overall accuracy	96.88% (LivDet2019)	Influence of finger pressure and duration on detection accuracy.
8	[40]	2020	Genetic algorithm optimized DenseNet for liveness detection	LivDet 2009, LivDet 2011, LivDet 2013, LivDet 2015	Accuracy	98.22% (mixed Livdet dataset)	Limited size of the dataset, requirement for specific sensor-matched datasets.
9	[41]	2020	Liveness detection using Circular Gabor Wavelet algorithm and SVM	-	Accuracy	99.968% (FAR)	Limited experimentation on artificial spoofing methods.
10	[42]	2020	Fingerprint and iris fusion-based liveness detection using statistical texture features	ATVS, LivDet2011	Precision, accuracy	94.7% (fingerprint detection), 97.8% (decision-level classification)	Applicability to certain types of attacks, dependency on dataset size.
11	[43]	2020	FLDNet CNN	LivDet 2013, 2015	ACE	1.76%	Improve accuracy on small fingerprints
12	[44]	2021	One-class Convolutional Auto encoder	-	D-EER	2.00%	Lack of generality
13	[45]	2021	Weighted MCNN	LivDet 2011, 2013, 2015, NUAA	Classification accuracy	2.42%	Diversity of sensors
14	[46]	2021	EaZy Learning	LivDet 2011, 2013, 2015	Accuracy	60.49%, 67.80%	Dependence on clustering
15	[47]	2021	Transformers + GANs	LivDet 2015	Accuracy	68.52% - 83.12%	Poor generalization
16	[48]	2021	Person-specific FPAD	-	Accuracy	100%	Limited generalization
17	[49]	2021	Multi-CNNs + Genetic Algorithm	Livedet datasets	Accuracy	+1.0%	Fixed-scale input limitation
18	[50]	2022	CNN	LivDet 2015	Accuracy	85.33%	Challenges with various materials
19	[51]	2022	Multi-filter Framework	LivDet 2009, 2011, 2013, 2015	ACC, ACE	99.15%, 0.85%	Extensive parameter configuration
20	[52]	2022	MFAS	24 subjects	Accuracy	100%	Limitations in real-life scenarios
21	[53]	2022	A-iLearn model for incremental learning	LivDet 2011, LivDet 2013, LivDet 2015	Overall Accuracy	Up to 49.57% improvement on new fake materials	Possibility of over fitting, need for further investigation of hand crafted and deep features
22	[54]	2022	Static-based approach with fusion of pores perspiration and texture features	LivDet 2013, LivDet 2015	Average Classification Error (ACE)	Biometrika: 0.11%, Italdata: 0.24%, Cross match: 0.21%	Limited amount of pore feature-based algorithms, difficulty in maintaining and updating algorithms
23	[55]	2023	Lightweight FLD network (LFLDNet) with CycleGAN and ResNet with MHSA	LivDet 2011, LivDet 2013, LivDet 2015	Average Classification Error	1.72 across all sensors, 95.27% accuracy on small-area fingerprints	Running speed affected by various factors, need for exploration of more effective FLD technology
24	[56]	2023	FPAD based on adversarial data augmentation and CNNs	LivDet2021	EER, BPCER, APCER, Liv. ACC	EER: 0.036, BPCER: 0.072, APCER: 0.000, Liv. ACC: 0.965	Limited dataset size, generalizability to other presentation attacks, possibility of adversarial attacks
25	[57]	2023	Fingerprint liveness detection using deep learning with LPDJH descriptor	LivDet 2009, LivDet 2011, LivDet 2013, LivDet 2015	Average EER	LivDet 2011: 3.95%	Difficulty in detecting low-resolution fingerprints, estimation difficulty due to limited test samples
26	[58]	2023	Fingerprint liveness detection utilizing CNNs	Socofing dataset	Accuracy, FAR, FRR	Accuracy: 98.964%, FAR: 0.215%, FRR: 7.251%	Imbalanced distribution in dataset, varying quality and characteristics of fingerprints, computational constraints

The Table 3 offers a comprehensive comparative analysis of various biometric cryptosystems, shedding light on different methods, datasets, metrics, accuracies, and limitations associated with these approaches. These biometric cryptosystems encompass a wide array of techniques such as fuzzy vault-based fingerprint cryptosystems, biometric key binding schemes, cryptographic authentication schemes based on discrete logarithm problems, and multi-biometric template security mechanisms, among others. Each method demonstrates specific strengths, including high accuracy, robust security, and effectiveness in protecting biometric templates. However, they also encounter challenges such as susceptibility to privacy attacks, computational complexity, limitations in dataset availability, and reliance on specific biometric modalities. Several future research directives emerge from this analysis. Firstly, there is a pressing need for the development of more efficient and secure biometric cryptosystems capable of addressing emerging security threats and vulnerabilities, including adversarial attacks, privacy breaches, and database-level attacks. Secondly,

research efforts should focus on enhancing the scalability and usability of biometric cryptosystems, particularly in real-world applications such as secure communication, access control, and identity verification. Thirdly, the exploration of novel cryptographic primitives, deep learning techniques, and hardware implementations could lead to innovative biometric cryptosystems with improved performance and reliability. Moreover, future research should prioritize the development of biometric cryptosystems that are compliant with regulatory requirements such as GDPR and HIPAA, while also ensuring user privacy and data protection. Additionally, efforts should be directed towards standardizing evaluation protocols and benchmarks to facilitate fair comparisons and reproducibility across different biometric cryptosystems. Furthermore, research should focus on the integration of biometric cryptosystems with emerging technologies such as blockchain and IoT to enhance security and interoperability in diverse application scenarios. By addressing these future research directives, the field of biometric cryptosystems can advance towards more secure, efficient, and user-centric solutions, thereby meeting the evolving needs of various domains including cyber security, healthcare, finance, and law enforcement.

Table 3: Comparative Analysis of Biometric Cryptosystems

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[59]	2015	Alignment-free fuzzy vault-based fingerprint cryptosystem using highly discriminative pair-polar (P-P) minutiae structures	FVC 2000 (DB1), FVC 2002 (DB1, DB2, DB3, DB4), FVC 2004 (DB2), FVC 2006 (DB2, DB3)	GAR, FAR	5.78%,0.06%	Security proof relies on complexity of brute force attack
2	[60]	2016	Feature level sequential fusion algorithm for biometric cryptosystems	Publicly available finger-vein database	FAR	1.47% FAR	Limited analysis of external security threats
3	[61]	2016	ECC-free biometric key binding scheme using Graph-based Hamming Embedding (GHE) and Minutia Vicinity Decomposition (MVD)	FVC 2002 (DB1, DB2)	GAR, FRR, FAR	GAR: 89%-97%, FRR: 3%-11%, FAR: 0.061%-0.16%	Vulnerable to privacy attacks like ARM and SKI, limited to matching fingerprint images of the same finger
4	[62]	2018	Biometric template security mechanism based on two-dimensional logistic sine map (2DLSM)	CASIA iris database, FVC 2002 (DB3)	Global shannon entropy	7.90	noise sensitivity, inability to generate biometric templates in real-time
5	[63]	2018	Secure cryptographic authentication scheme based on discrete logarithm problem	CASIA iris database	Inner and outer Hamming distance distributions, FA, FR, BER	1.16%, 28.3%	need for enhancing iris codes for better performance
6	[64]	2019	Biometric-based cryptographic key generation mechanism using convolution coding principles (BioKEY)	MIAS, FVC2002, FVC2004	True positive rate	95.12%	struggle with noisy or blurred fingerprints, significant computational time
7	[65]	2019	ECC-based mutual authentication scheme for Smart Grid communications using biometric approach	Inhouse Dataset	Computation cost	8.92 ms	vulnerability to certain attacks
8	[66]	2020	Fuzzy vault method for template security of multimodal biometric systems with face and fingerprint images	Virtual face and fingerprint database	GAR	99%	balancing security and accuracy
9	[67]	2020	Enhanced iris recognition approach using hyperelliptic curve cryptography (HECC)	CASIA Iris V-4, IITD iris datasets	Accuracy	99.74%	limitations with fuzzy extractor's parameters and potential attacks
10	[68]	2020	Multibiometric cryptosystem for user authentication	100 subjects with iris and fingerprint modalities	FRR, EER	0.01, 0.005	Limitations in accuracy and precision due to variability of biometric data
11	[69]	2020	Multi-biometric template security method based on unique graph generation	CIE Fingerprints of IITD Database	ERR	Low ERR of 0.66%	Dependency on input sample quality
12	[71]	2020	Multimodal biometric cryptosystem using fingerprint and ear	Fingerprint and ear image datasets	accuracy	98.76%	Reliance on systems for high accuracy
13	[72]	2021	Biometric cryptosystem based on random projection and back propagation neural network	NIST SD4, Faces94	FNMR	2.90%	Time-consuming training process for off-line enhancement, need for further research on multiple-biometric template protection
14	[73]	2021	Fingerprint biometric cryptosystem based on fuzzy commitment scheme and CNN	FVC2000 DB2-A fingerprint database	FAR, FRR, EER	FAR: 1.25%, FRR: 1.15%, EER: 2.83%	Challenge of determining reference point precisely, need for larger training set
15	[74]	2021	Cancellable biometric authentication framework leveraging GA	ORL, FERET, LFW	AROC	09998	Need for further validation on larger and more diverse datasets
16	[75]	2021	Cancellable biometric security system based on advanced chaotic maps	In House	EER	0.593	Non-invertibility of biometric transformations, minor changes in biometrics affect hash functions
17	[76]	2021	Cancellable Biometrics Vault (CBV) using chaffing and winnowing	CASIA-V3-Interval	FRR, FAR	6.92%, 0.001%	unsuitable for real-time applications due to computational complexity
18	[77]	2021	Multi-biometric cryptosystem based on Modulus Fuzzy Vault algorithm	IIT Delhi Iris Database, IIT Delhi Palmprint Database, IIT Delhi Ear Database, NIST Special	ROC curve, FAR, FRR, GAR	0.96	Limited focus on specific algorithms, need for experiments with natural image databases

19	[78]	2022	Machine vision gait-based biometric cryptosystem using fuzzy commitment scheme	CMU MoBo, CASIA A	FAR, FRR	0%	Recognition limitations due to variations like complex background and occlusions
20	[79]	2022	Asymmetric cryptosystem based on elliptic curve algorithm and optical scanning cryptography (OSC)	Inhouse	Information entropy	7.95	Speed limitations in encryption and decryption, vulnerability to cipher-text-only attacks
21	[80]	2022	Fuzzy extractor for generating cryptographically strong keys from biometric images using deep learning and code-based cryptosystems	LFW, CelebA	Accuracy	93%	Requirement for large storage capacities with deep learning, susceptibility to attacks based on machine learning model inversion
22	[81]	2022	Block chain-based user re-enrolment scheme for biometric-based authentication systems	Artificial dataset	Time complexity	0.1515	Computational complexity dependent on number of users, assumption of non-adversarial participants
23	[82]	2022	Secured multi-biometric template protection using Lagrange's interpolation	The Hong Kong Polytechnic University dataset	Accuracy	99.9816%	Increase in database size, may not meet all ISO/IEC 24745 requirements
24	[83]	2023	Hardware design of secure cancellable biometric cryptosystem based on 3D chaotic map	ORL, FVC, LFW	EER, AROC	EER of 6.2460×10^{13}, AROC=0.9998	Needs further study on different types of attacks and larger datasets
25	[84]	2023	Cancelable biometric authentication mechanism using 3D chaotic maps, PWLCM, logistic map, and DNA sequencing theory	Various face and palm print datasets	AROC, FAR, DH, SSIM, PSNR	AROC = 1, AFAR of 6.2×10^{3}, ADH=0.8755, PSNR=8.2061	Increase in computational resources required
26	[85]	2023	Enhanced Biometric Cryptosystem (BCS) using ear and iris modalities based on BRIEF	AMI, UBIPr	NRMSE	1.7486	Susceptibility to database-level attacks, need for more sophisticated security techniques
27	[86]	2023	Multi-biometric secure-storage scheme based on deep learning and crypto-mapping techniques	CASIA V4, MICHE, ICUB, MobiFace	AUCROC, EER	0.054	Need for hardware implementation, possibility of over fitting, trade-off between security and recognition performance
28	[87]	2023	Biometric key generation and multi-round AES cryptosystem for improved security	Inhouse	Energy efficiency	24.67 ms	Lack of dataset details, limited encryption techniques tested, limited performance metrics reported

The Table 4 presents a comparative analysis of cancellable biometric methods, offering insights into various techniques, datasets, metrics, accuracies, and limitations associated with these approaches. These methods encompass a wide range of strategies such as protection methods for fingerprint templates, generating masterprints for impersonation, one-factor cancellable biometric authentication schemes, secure triplet loss for training deep learning models, and multi-server authentication using cancellable biometrics and PUF, among others. Each method exhibits specific strengths, including high recognition accuracy, robustness against attacks, and effectiveness in protecting biometric templates. However, they also face challenges such as computational complexity, vulnerability to specific attacks, dependence on dataset characteristics, and limitations in adaptability to diverse biometric modalities. Several future research directions emerge from this analysis. Firstly, there is a need for the development of more efficient and scalable cancellable biometric techniques, particularly in addressing computational complexity issues and improving performance across diverse biometric datasets. Secondly, research efforts should focus on enhancing the security and robustness of cancellable biometric methods against evolving threats and sophisticated attacks, including template inversion attacks, brute-force attacks, and adversarial manipulations. Thirdly, the exploration of novel approaches such as deep learning-based fusion techniques, chaotic-based cancellable systems, and image style transfer for biometric authentication could lead to innovative solutions with improved performance and reliability. Moreover, there is a need for standardized evaluation protocols and benchmarks to facilitate fair comparisons and reproducibility across different cancellable biometric methods. Additionally, future research should prioritize the development of cancellable biometric techniques that are user-friendly, privacy-preserving, and compliant with regulatory requirements such as GDPR and HIPAA. Furthermore, efforts should be directed towards investigating the usability and acceptability of cancellable biometric systems in real-world applications such as access control, identity verification, and secure authentication. By addressing these future research directives, the field of cancellable biometric methods can advance towards more secure, efficient, and user-centric solutions, thereby meeting the evolving needs of various domains including cyber security, healthcare, finance, and law enforcement.

Table 4: Comparative Analysis of Cancellable Biometric Methods

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[88]	2017	Protection method for fingerprint templates using fused structures at feature level	FVC 2002 and 2004 databases	EER, separability, KS test	DB1: 2.19%, DB2: 1.6%, DB3: 6.14%; DB1: 11.89%, DB2: 12.71%, DB3: 17.6%	Computational cost for large-scale databases, non-invertibility
2	[89]	2017	Generating "MasterPrints" for impersonation in partial fingerprint-based authentication systems	FingerPass DB7 dataset, FVC2002 DB1-A dataset	Imposter Match Rate (IMR)	6.77%, 1.31%,0.36%, 3.51%	Dataset imbalance, applicability limited to minutiae-based systems

3	[90]	2018	One-factor cancellable biometric authentication scheme using Indexing First Order hashing	FVC 2002 and FVC 2004	EER, genuine-imposter distribution	4.26%, 2.67%	Vulnerable to COA, KPA, CCA; identifier unlinkability
4	[91]	2019	One-factor cancellable palmprint biometric recognition scheme based on OIOM hash and MSH	PolyU and TJU palmprint databases	Recognition accuracy	98.07%	No explicit mention of limitations
5	[92]	2020	Universal solution for multi-biometric systems using deep neural networks	IITD Iris and MMU2	DI, EER	DI: 10.35, EER: 0.12	Limited adaptiveness, need for sensitive environment
6	[93]	2020	Secure Triplet Loss for training end-to-end deep learning models for cancellable biometric templates	Off-the-person ECG and unconstrained face images	EER	12.56%, 13.99%	Vulnerability under specific attack scenarios, impact of variability factors
7	[94]	2020	Constrained Optimized Similarity-based Attack (CSA) on cancellable biometrics using IoM hashing	LFW dataset	Success rate, FAI rate, TAI rate	99.19%, 98.58%, 0.2866%	Dependency on protected template information
8	[95]	2020	Non-invertible cancellable fingerprint template based on Delaunay triangulation	FVC2002 database	EER	1.6%, 2.5%, 2.8%, 3.3%, 2.4%	Concerns about compromised acquisition device, need for larger dataset
9	[96]	2021	Highly optimized user template construction for fingerprint-based authentication	Nine different fingerprint databases	EER	0%	Performance limited by image quality, suggestion for multi-modal system
10	[97]	2021	Constrained-optimized similarity-based attack on cancellable biometrics using IoM hashing and BioHashing	LFW dataset	SAR, FAI	72.74%, 72.11%	Fixed model complexity, over fitting reliance on IoM hash code size
11	[98]	2021	Cancellable template using GCD	Facial, fingerprint, iris, palm print	EER, AROC	High AROC up to 99.59%, low EER down to 0.04%	High-quality initial biometric images needed
12	[99]	2021	Feature-adaptive random projection	FVC2002 DB1-DB3, FVC2004 DB2	EER, GAR, FAR	1.0%, 2.0%, 4.0%, 11.0%	Need for more discriminatory feature descriptor
13	[100]	2021	Multi-server authentication using cancellable biometrics and PUF	LFW dataset	CMC curve, ROC curve, DIR curve, DET curve	$18t_h$	Lack of evaluation against sophisticated attacks
14	[101]	2021	BioCanCrypto: biocryptosystem on fingerprint cancellable templates	FVC2002 (DB1, DB2, DB3)	EER, FRR	0.12%, 0.59%, 2.71%	Limited exploration in different feature spaces
15	[102]	2021	Watermarking reinforcement scheme	BioSecure, FVC2002 DB1	FMR, GMR, EER	1.98%	Limited to single biometric trait
16	[103]	2022	Absolute Value Equations Transform (AVET)	Eight datasets for various biometrics	GAR	93.22%, 91.91%, 96.2%	Fixed sample size, vulnerability to brute-force attacks
17	[104]	2022	Multi-biometric cancellable scheme using deep fusion and deep dream	Nine images from each biometric modality	NPCR, PSNR, SSIM, UIQ, SD, UACI	99.158%, 24.523, 0.079, 0.909, 59.582, 23.627	Computationally intensive, requires enrollment of all seven images
18	[105]	2022	Cancellable multi-biometric identification scheme using ACM	FVC2002, ICE 2005	EER, ROC curve	0.0005, 0.0019	Lack of large-scale database, need for reliable parameter estimation
19	[106]	2022	Cancellable SoftmaxOut Fusion Network (CSMoFN)	AR, Ethnic, Facescrub, IMDB, Wiki, YTF	EER	6.67%, 6.71%	Risk of CB template inversion
20	[107]	2022	Selective encryption and deep learning-based fusion technology	"Lena" image, face images	Correlation, entropy, ROC curve, AROC	0.0008, 0.0019	Limited size of the dataset
21	[108]	2023	Deep learning and style transfer for cancellable biometric system	1800 images dataset segmented into face and fingerprint biometrics	NCC, MSE, PSNR, SSIM, UIQ, SD, UBER	Average NPCR: 99.26, PSNR: 23.28, SSIM: 0.0405, UIQ: 0.7492, SD: 60.442, UACI: 24.268	Need for consistent physical condition, reduced reliability in certain environmental conditions
22	[109]	2023	Multi-Biometric Feature Hashing (MBFH)	Retina, finger veins, palm, dorsal vein images	Hamming, Spearman, Jaccard pairwise distances	Average values over 0.9 in Spearman, Jaccard, and hamming distances	Applicability to adaptable and featherless biometric features, need for adaptability for adding white Gaussian noise
23	[110]	2023	Biometric template protection scheme for Euclidean and Cosine metrics	AR Face, CASIA FaceV5, ORL, LFW	EER	8.559, 0.154	Not optimal for other types of biometric authentication systems, slightly higher computational costs due to pre-processing
24	[111]	2023	Chaotic-based cancelable face recognition system using convolution kernels	AT&T, YALE, UFI, LFW, FERET	Accuracy	98.43%	Trade-off between system performance and user privacy, need for adaptation to environmental circumstances
25	[112]	2023	Biometric authentication system using image style transfer	Face image database with key images	Correlation coefficients, ROC curves	AUC values > 0.9 in most cases	Problem of key image similarity, need for restrictions in setting key images in future studies

The presented Table 5 offers a comprehensive analysis of various general-purpose user authentication methods, covering a wide range of techniques, datasets, metrics, accuracies, and limitations. These methods encompass diverse approaches such as orientation extraction, fingerprint reconstruction, fake biometric trait detection, local model-based classification, partial fingerprint matching, minutiae extraction evaluation, and multi-modal biometric authentication,

among others. Each method demonstrates specific strengths, including high accuracy rates, robustness against attacks, and efficiency in authentication processes. However, they also encounter challenges such as dataset limitations, computational resource requirements, dependence on specific features, and vulnerabilities to spoofing attacks. Several future research directions can be identified from this analysis. Firstly, there is a need for the development of more robust and secure authentication methods, particularly in addressing vulnerabilities to spoofing attacks and enhancing resistance against adversarial manipulations. Secondly, research efforts should focus on improving the scalability and efficiency of authentication systems, especially in handling large-scale datasets and reducing computational resource requirements. Thirdly, the exploration of novel biometric features, fusion techniques, and machine learning algorithms could further enhance the accuracy and reliability of authentication methods across diverse modalities. Moreover, the development of privacy-preserving authentication techniques and the investigation of human-centric design principles could ensure the usability and acceptability of authentication systems in real-world scenarios. Lastly, efforts should be directed towards standardizing evaluation protocols and benchmarks to facilitate fair comparisons and reproducibility across different authentication methods. By addressing these future research directives, the field of general-purpose user authentication can advance towards more secure, efficient, and user-friendly authentication systems, thereby meeting the evolving needs of various applications such as cybersecurity, access control, and identity verification.

Table 5: Comparative Analysis of General Purpose User Authentication Methods

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[113]	2011	Orientation Extraction	FVC2006 DB2	Average Error	0.206%	Limited benchmark dataset
2	[114]	2012	Fingerprint Reconstruction	FVC2002 DB1A, DB2A	Successful Match Rate	86.48%, 86.96%	Performance limitations for certain scenarios
3	[115]	2013	Fake Biometric Trait Detection	Fingerprint, Iris, 2-D Face	Classification Error Rate	<3%	Computational resources, full image access
4	[116]	2015	Local Model-based Classification	FVC 2000, 2002, 2004	Accuracy	96.7%, 96.5%	Allocation of low-quality fingerprints for testing
5	[117]	2016	Partial Fingerprint Matching	FVC2000, FVC2002	EER	1.17%, 1.4%	Dependence on enhancement algorithm
6	[118]	2017	Minutiae Extraction Evaluation	FVC 2002, FVC2004, Synthetic data	Average positional error, Average orientation error	3.48%, 0.06%	Environmental variations
7	[119]	2017	GLDM for Bio-Cryptosystems	PUCPR, GPDS-300	Classification Error Rate	7%, 17%	Limited positive signature samples
8	[120]	2018	Pore Extraction using CNN	Touch-based, Touchless, Latent	Detection Rate, False Alarm Rate	84.69%, 15.31%	Need for high-resolution images
9	[2]	2019	Pore Comparison	FVC2002 DB I, II	EER, FMR1000s	1.86%, 0.12%	Dependence on alignment accuracy
10	[121]	2020	Fingerprint Enhancement & Reconstruction	FVC2002, FVC2004	TAR	97.95%, 94.09%	Impact of dirt, age, medical factors
11	[122]	2020	DeepPoreID Matching	DBI, DBII	EER, FMR100	35% increase in EER	Dependence on large quantity of sweat pores
12	[123]	2020	Homomorphic Encryption	FVC2002 DB2	EER	9.23% EER	Time-consuming encryption
13	[124]	2020	SSO Feature Extraction	150 images	FAR, FRR, CVR	FAR: 0.00, FRR: 0.00666, CVR: 99.334%	Small dataset size
14	[125]	2020	FHE-based Authentication	NIST SD9	EER	0.053%	Time-consuming encryption
15	[126]	2020	BioSec Framework	FVC-2004 DB3	Accuracy	70%	Vulnerability of symmetric encryption
16	[127]	2020	Reversible Watermarking	FVC2002	PSNR, Matching Score	99.61%	Sensitivity to noise and compression
17	[128]	2021	Ensemble Matching	NIST SD27, GCDB, MOLF DB1	Rank-1 identification rate	74.03%	Choice of supervised classifier
18	[129]	2021	DRUnet Segmentation	FVC Databases	Jaccard similarity, Dice score	97.21%, 94.73%	Small training dataset
19	[130]	2021	MiDeCon Quality Assessment	FVC 2006	FMR	10^{-1},	Complex procedure
20	[131]	2021	ASRA Image Alignment	FVC2002, FVC2004	Authentication accuracy	99%	Dependence on ROI extraction
21	[132]	2021	VSS & Super-resolution	FVC2002 DB1	Genuine match rate	94%,	Need for higher quality fingerprint images
22	[133]	2021	Dual-filter Framework	LivDet 2013, 2015	Accuracy	97.56%	Framework agnostic to filter type
23	[134]	2021	Multi-modal Biometric Authentication	228 image	signal data & EER, ROC curves	Classification rate of 100%	Not tested on large scale dataset
24	[135]	2022	Secure Authentication	FVC2002 DB1-DB3, FVC2004 DB1-DB3	EER	0.99%, 3.28%	Balance between security and performance
25	[136]	2022	Embedded System Authentication	In-house	Accuracy	-	Robustness against attacks
26	[137]	2022	Gabor Filter & CNN	FVC2006	Validation accuracy	99.33%	Management of smaller training sets
27	[138]	2022	Deep CNN Matching	In-house	Training & validation accuracy	100%	Lack of diversity in dataset
28	[139]	2022	Multi-factor Authentication	CASIA-FingerprintV5, FVC2002 DB1	FAR	0.81%	Limited size of dataset

29	[140]	2023	Brute-force Attack	In-house	FAR	4.1410^6	Need for smartphones to support sensor hot plugging
30	[141]	2023	Joint Authentication & Spoof Detection	FVC 2006 DB2A, LiveDet 2015	TAR, FAR, ACE	TAR = 100%, ACE = 1.44%	Need for larger and more diverse datasets
31	[142]	2023	Homomorphic Encryption & ML	PolyU, SOKOTO	F1-score	PolyU: 93.6%, SOKOTO: 98.2%	HE-based computation challenges
32	[143]	2023	Multi-factor Authentication	In-house	System Usability Scale	90%	Small sample size
33	[144]	2023	AVAO enabled DMN	CASIA Fingerprint Image Database	Accuracy, Sensitivity, Specificity	0.927, 0.938, 0.930	Need for robust pre-processing

The Table 6 offers a comprehensive comparative analysis of various fingerprint indexing methods, encompassing different approaches, datasets, metrics, accuracies, and limitations. Notable methodologies include distributed frameworks, GPU-based systems, absolute registration approaches, minutiae-based algorithms, deep learning-based models, and privacy-preserving indexing techniques. These methods exhibit diverse strengths, such as high processing speeds, improved accuracies, and workload reduction, while also facing challenges like hardware/software dependencies, scalability issues, and computational costs. Several future research directions emerge from this analysis. Firstly, there's a need for continued exploration of deep learning techniques in fingerprint indexing, especially in improving accuracy and reducing computational costs. Additionally, addressing scalability concerns and hardware dependencies would be crucial for real-world deployment. Secondly, the development of privacy-preserving indexing methods is vital to ensure the security and confidentiality of biometric data, especially in light of increasing concerns regarding data privacy. Thirdly, research efforts should focus on enhancing the robustness and generalization capabilities of indexing algorithms, particularly in handling diverse datasets with varying quality and characteristics. Moreover, investigating novel features and descriptors, as well as exploring multimodal approaches, could further improve indexing performance and reliability across different biometric modalities. Lastly, efforts should be directed towards standardizing evaluation protocols and benchmarks to facilitate fair comparisons and reproducibility across different indexing methods.

Table 6: Comparative Analysis of Fingerprint Indexing Methods

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[145]	2014	Two-level distributed framework for fingerprint matching	Large synthetic database created with SFinGe, NIST databases	Spped up	312.8684 sec	Distribution of synthetic fingerprint dataset, hardware resources, scalability of matching algorithms
2	[146]	2015	Fingerprint identification system using GPUs	FVC2002, DB14	EER	2%	Potential benefits of GPU-based algorithms, need for a two-stage system for practical applications
3	[147]	2016	Absolute registration approach with pose constraint	Public domain fingerprint databases, NIST Special Database 14, database of 1,000,000 rolled fingerprints	Error Rate	2.33%	High dimensionality of the Minutia Cylinder-Code (MCC) descriptor, need for further research on noise and variability
4	[148]	2017	Fingerprint indexing algorithm based on minutiae pairs and convex core point	FVC2000DB2A+B, FVC2000DB3A+B, FVC2002DB1A+B, FVC2004DB1A+B, NIST's DB4 and DB14	TPR	2.18	Lack of detail regarding parameters, hardware/software differences
5	[149]	2019	Minutiae-centred fingerprint indexing method with deep convolutional neural network (DCNN)	FVC2002 DB2a, FVC2002 DB2b, FVC2004 DB1a, NIST special database 4, NIST special database 14	Error rate, penetration rate	0.25%, 10%	High computational cost for MCC-based method
6	[150]	2019	DeepPrint: Deep network for learning fixed-length fingerprint representations	NIST SD4, last 2,700 pairs of NIST SD14	Rank-1 search accuracy	98.80%	Fixed length representation, Saturation of existing benchmarks
7	[151]	2020	CNN-based Combination of Nearest Arrangement Indexing (CNNAI)	FVC2004, NIST SD27 latent fingerprint databases	Rank-1 identification rate	80%, 85.4%	Dependence on the availability of minutiae points, failure of MINU-EXTRACTNET
8	[152]	2020	Pore-based indexing method for high-resolution fingerprints	Hong Kong PolyU high-resolution fingerprint databases (DBI and DBII), IITI-HRFP, IITI-HRF	Pre-selection error rate, penetration rate	67%, 49%, 42%,28%, 10%	Effectiveness limited on databases of high-resolution full fingerprints
9	[153]	2020	Latent fingerprint indexing using minutiae based rotational and translational features	IIIT-D latent fingerprint dataset	Accuracy, indexing time	Accuracy up to 94.77%, average indexing time ranging from 5.29 to 25.92 seconds	Challenges with partial impression, background noise, poor ridge clarity, large non-linear distortions
10	[154]	2021	Indexing algorithm based on clustering of minutia cylinder codes (MCC)	Public data, proprietary data, synthetic data, NIST SD4, NIST SD14	Error rate, Penetration Rate	53.9%, 0.01%	Reduced search space, lower performance on databases with different characteristics or quality levels, time complexity
11	[155]	2021	Pore-based indexing and refinement	DBI, DBII	Accuracy	95.16%	Computation increase with pores

12	[156]	2021	Inverted index for minutiae-based search	FVC2002 DB1a, Private dataset	Error rate	0.42%	High search rates
13	[157]	2021	PalmHashNet for palm print indexing	CASIA, IIT Delhi, Tongji, PolyU II	Accuracy	Above 99% accuracy	Not tested on large datasets
14	[158]	2021	MDSH-based indexing for multi biometric retrieval	IIT Delhi (iris, face, palm print)	Hit rate, Penetration rate	68%, 100%	Computational costs
15	[159]	2022	Quality score-based search for biometric identification	FERET, CASIA, FVC2002	Workload	Up to 38% reduction	High intra-class variability
16	[160]	2022	FKPIndexNet for finger-knuckle-print identification	PolyU-FKP, IITD-FKP	Error rate, Penetration rate	0.32%, 100%	Limitations in index table generation
17	[161]	2022	Signal-based fusion for biometric database indexing	FERET, FRGCv2, IJB-A	Workload	Reduced workload by 30%	Requires high-quality images
18	[162]	2023	Privacy-preserving multi-biometric indexing	Composite dataset	Workload	Workload reduction by 57%	Use of composite dataset
19	[163]	2023	Attention-Driven Fingerprint Recognition Network	NIST SD4, NIST SD14	Accuracy	99.93%	Not using domain knowledge
20	[164]	2023	Gravitational Search Decision Forest for fingerprint recognition	NIST SD27, FVC2004	Precision, Recall, F-measure	93.90%, 96.25%, 95.06%, 96.25%	Recognition rate optimization

The Table 7 presents a comprehensive overview of various methodologies employed for infant and child biometric recognition, highlighting the diverse approaches, datasets, metrics, accuracies, and limitations associated with each method. Several studies focus on fingerprint recognition, utilizing techniques ranging from image enhancement and CNN-based ridge flow estimation to pre-trained CNN models. These approaches demonstrate promising accuracies, with [166] achieving up to 99.5% True Accept Rate (TAR) for infant fingerprints. However, challenges persist, such as the lack of automated alignment methods and issues with small fingerprint sizes and low image quality, as noted by multiple authors. Moreover, alternative biometric modalities like footprints, iris, and outer ear shape are explored, showing potential for high accuracy, as evidenced by [168] 99.30% accuracy in footprint identification and [170] low Equal Error Rate (EER) for outer ear shape recognition. Nonetheless, these methods encounter hurdles related to data collection challenges, limited dataset sizes, and transparency in processes and facilities. Future research in infant and child biometric recognition could address these limitations by focusing on several key areas. Firstly, efforts should be directed towards developing automated alignment techniques to enhance accuracy and efficiency, especially for infant fingerprints. Secondly, there is a need for larger and more diverse datasets to ensure robust model performance and generalization. Thirdly, exploration of multi-modal biometric systems, as suggested by [172], could provide enhanced accuracy and reliability by combining different biometric traits . Lastly, advancements in image acquisition technologies and computing resources could facilitate the collection of high-quality images and alleviate challenges associated with low image quality and computing resource requirements, as highlighted by [169]. By addressing these research directions, future studies can contribute towards the development of more robust, accurate, and reliable infant and child biometric recognition systems, thereby enhancing their applicability in various domains such as healthcare, security, and childcare.

Table 7: Comparative Analysis of Infant and Child Biometric Recognition Methods

Sl No	Paper	Year	Method	Dataset	Metric	Accuracy	Limitation
1	[165]	2016	Image enhancement, CNN-based ridge flow estimation	Infant fingerprint database	Rank-1 identification accuracy	73.98%	Lack of automated fingerprint alignment for infants
2	[166]	2016	Evaluation of child fingerprint recognition using NFIQ 2.0 metric	Dataset of infants aged 0-5 years	True Accept Rate (TAR), False Accept Rate (FAR)	TAR: 99.5% (500 PPI), 98.9% (1270 PPI)	Piecewise linear model interpretation, dataset resolution variations
3	[167]	2021	InfantPrints system with custom-built fingerprint reader	Longitudinal infant fingerprint database	TAR, FAR	95.2%, 1.0%	Lack of automated alignment method
4	[168]	2021	Footprint identification using wavelet feature extraction and K-NN	Dataset of 30 babies with footprint images	Accuracy, Precision, Recall	99.30% accuracy, 90.17% precision, 89.44% recall	Small dataset size, single feature extraction method
5	[169]	2022	Neonate fingerprint classification using deep learning	Dataset with manual labelling of fingerprint images	Accuracy	78.4%	Computing resource requirements, challenges in collecting high-quality images
6	[170]	2023	Biometric recognition systems for infants (fingerprint, iris, outer ear shape)	Dataset from volunteer infants and children	Equal Error Rate (EER), Failure to Acquire (FTA)	EER: 0.16% (outer ear shape)	Challenges in data collection, limitations in facility and process transparency
7	[171]	2023	Multi-instance contingent fusion technique for infant fingerprint verification	Dataset of infant fingerprints	Verification accuracy	73.8% (1 month interval), 57.14% (6 months interval)	Absence of ancillary information, greater time interval necessitates fusion
8	[172]	2023	Multi-modal Biometric Recognition for Toddlers and Pre School Children (M^2BRTPC)	Dataset of iris and fingerprint modalities from children aged 18 months to 4 years	Accuracy	96.3%	Restricted age range, improvements over existing methods without claiming absolute accuracy
9	[173]	2023	Pre-trained CNN model for newborn and toddler fingerprint recognition	Dataset of fingerprint images from 154 subjects	Accuracy	82.47%	Small size of fingerprints, low image quality

10	[174]	2023	CNN-based children recognition system using contact less fingerprints	CLCF dataset for experimentation	Accuracy, Equal Error Rate (EER)	Accuracy: 98.46%, EER: 1.99%	Need for larger datasets, exploration of different enhancement and extraction techniques

Several important conclusions can be drawn from a comprehensive analysis of the literature on latent fingerprint recognition and biometric identification systems. With a wide range of strategies being investigated, such as deep learning models, hybrid approaches, and sophisticated algorithmic techniques, these systems have undergone tremendous evolution. Due to this diversity, difficult fingerprint identification problems like partial prints, noise, and inconsistent quality have been significantly improved. The literature does, however, also highlight enduring difficulties. The comparison of various methods is made more difficult by the absence of standard evaluation methodologies, and algorithmic bias and generalizability are raised by the small size and lack of diversity of the datasets that are currently accessible. Moreover, a lot of sophisticated techniques, especially those centred on deep learning, have a high computational complexity, which makes real-time implementation challenging. Future work in this area has to concentrate on creating uniform benchmarks, growing datasets, maximizing computational effectiveness, and resolving moral issues with bias and privacy. Advantage, disadvantage, area of improvement and ethical implication of the fingerprint biometric studied in the manuscript are discussed below.

- **Advantage:**
  - **Diverse Methodologies:** The literature study demonstrates the wide variety of approaches, such as hybrid approaches, descriptor-based tactics, deep learning models, and different algorithmic advancements, that are employed in fingerprint recognition and biometric systems. Due to the variety of methods available, researchers are able to tackle several facets of latent fingerprint recognition issues, including handling partial prints, noise, and quality differences. For instance, descriptor-based methods may provide computational efficiency, while deep learning models might enhance accuracy by discovering complex patterns in fingerprint data. The use of hybrid techniques, which blend conventional and cutting-edge methodologies, shows an effort to capitalize on each approach's advantages and advance latent fingerprint recognition technology.
  - **Quantifiable Performance Metrics:** Numerous studies provide comprehensive performance indicators, like F1 scores, recall, precision, and rank-1 accuracy, which are essential for evaluating the efficacy of various techniques. Benchmarking these measures against popular datasets such as FVC2004 and NIST SD27 is common practice. Researchers and practitioners find it easier to compare the effectiveness of different techniques, evaluate advances, and pinpoint areas that require additional development when such measurements are transparently reported. Because of its transparency, the community is able to monitor advancements and set performance standards for systems to come.
  - **Cutting-edge Techniques:** Numerous novel methods have been presented, such as deep learning-based minutiae extraction, deformable minutiae clustering, and data augmentation using Generative Adversarial Networks (GAN). Especially for under-represented instances, GAN-based augmentation might artificially generate extra training data, hence improving the system's resilience. Latent fingerprint identification systems are becoming more accurate and scalable thanks to deformable minutiae clustering and deep learning techniques. This makes them more suitable for use in forensic applications, particularly in challenging situations where the prints are imperfect or deteriorated.
  - **Real-world Testing:** The fact that numerous suggested techniques are evaluated on a range of databases and datasets, including NIST SD27 and FVC2004, is a noteworthy strength of the examined literature. This makes it possible to test the algorithms in a variety of scenarios, simulating the unpredictability of the actual world. By doing this, scientists may evaluate how reliable and broadly applicable their methods are, making sure they work effectively outside of carefully regulated lab environments. For technology meant to be used in actual applications, including forensic investigations and security systems, this real-world applicability is essential.
- **Disadvantages:**
  - **Lack of Uniformity in Evaluation Standards:** Although a large number of research offer performance indicators, there is a notable deficiency in uniform assessment procedures between various inquiries. Direct comparisons between approaches are challenging because to variations in benchmarking procedures, quality criteria, and dataset sizes. Because of this, some accuracy statistics might be inflated, which makes it more difficult to determine whether strategy is better in an unbiased manner. A technique that works well on a small, clean dataset, for instance, might not translate as well to larger, messier real-world data. Additionally, because there isn't a single benchmark, reported accuracy gains may vary depending on the context and may not apply to different circumstances.
  - **Dataset Limitations:** The small size and lack of diversity of datasets used for testing and training is a problem that keeps coming up in the literature. A lot of research papers admit that their datasets are too tiny or homogeneous to adequately capture the heterogeneity present in real-world circumstances, especially when it comes to latent print datasets. This suggests that the models may be over fitting to the unique features of these small datasets, which raises questions about their generalizability. A system that works well on the dataset it was trained on but has trouble with under-represented populations or situations, including various skin kinds, lighting settings, or sensor types, may also be introduced by a lack of diversity.

- **Computational Complexity:** There is a drawback to several of the sophisticated methods, especially those that depend on deep learning: higher computational demands. These techniques frequently call for additional processing power, memory, and time, but they may provide considerable accuracy gains. It is difficult to apply these techniques in real-time scenarios, such as live fingerprint recognition at security checkpoints or mobile devices with constrained computing power, due to the rise in computational complexity. For example, deep learning models can yield higher recognition rates, but their resource needs make them impractical for use in real-time, low-power devices.

- **Areas for Improvement:**

  - **Standardized Benchmarks and Evaluation Protocols:** The creation of uniform protocols for comparing fingerprint recognition algorithms across various datasets is among the most urgent need. This would entail defining testing settings, dataset properties, and consistent evaluation criteria. Researchers may more precisely compare the performance of various approaches when a consistent framework is in place, which helps them determine which techniques are actually improving the state of the art. Exaggerated performance claims might be lessened by using a consistent evaluation process, which would guarantee that all approaches are examined under comparable circumstances [118].

  - **Larger and More Diverse Datasets:** Future study should concentrate on increasing the amount and diversity of latent fingerprint databases in order to solve the limitations of the present investigations. This would include gathering information from a larger variety of sources and in a greater variety of settings, including various skin tones, ambient variables, and sensor kinds. By using larger, more representative datasets for algorithm training, researchers can improve system robustness and minimize bias. It would be easier to make sure that the algorithms can handle the whole range of real-world issues if more diverse latent prints were included, such as those with different levels of noise, distortion, and completeness [145].

  - **Improving Computational Efficiency:** Because many of the methods are computationally demanding, optimization is required to shorten processing times without compromising accuracy. For real-time applications, where speed is a crucial component, this is especially crucial. To lessen their processing footprint, researchers can concentrate on creating lightweight algorithms or improving already-existing deep learning systems. For example, the models could be deployed in resource-constrained contexts like mobile devices or edge computing platforms if they are optimized to perform well on GPUs or embedded systems [25].

  - **Integration with Other Biometric Modalities:** Integrating fingerprint identification systems with other biometric modalities, such facial recognition or iris scanning, is another possible avenue for advancement. When one modality is impaired (a partial fingerprint, for example), this multi-modal method may enhance overall accuracy and resilience. Combining information from several biometric sources may also improve security by making it more difficult for hackers to impersonate the system. Future developments may proceed in the direction of sensor fusion technique research and the creation of reliable multi-modal biometric authentication algorithms [42].

  - **Ethical Considerations and Privacy:** Future research must also address ethical issues including privacy concerns, data protection, and algorithmic bias given the growing usage of biometric technologies in sensitive applications like security and forensics. Strong legal frameworks are especially necessary for forensic applications in order to guarantee the validity and acceptability of the data gathered using these systems. Researchers should investigate the moral ramifications of gathering and using biometric data, making sure privacy laws like GDPR are followed, and creating clear procedures for handling user consent and data [180].

- **Ethical Implications:**

  - **Privacy Concerns:** Fingerprint biometric systems create serious privacy concerns since they rely on the gathering, storing, and processing of private information. Biometric information, like fingerprints, is unchangeable after it is compromised, unlike passwords or other identity-related information. Biometric data is particularly susceptible to exploitation because of this. Identity theft, unapproved tracking, or monitoring might result from unauthorized access to fingerprint databases or data breaches. These worries are heightened by the collecting of biometric data, frequently without users' express or informed agreement. The collection, storage, and sharing of this data must adhere to strict processes in order to comply with data protection legislation, such as the General Data Protection Regulation (GDPR) in Europe. Crucial to ethical compliance are letting users know how their biometric data is being used and providing them with the choice to opt out. To reduce these concerns, biometric systems should place a high priority on openness, safe encryption, and privacy-preserving methods like homomorphic encryption or cancelable biometrics.

  - **Algorithmic Biases:** The possibility of algorithmic bias in fingerprint biometric technologies presents another ethical dilemma. Research has indicated that the accuracy of biometric systems can vary based on the demographic attributes of the people being scanned, including age, gender, and skin tone. For instance, under representation of particular populations in training datasets may result in reduced accuracy for those groups. Particularly in forensic or security scenarios, where a false match or inability to identify a legitimate individual can have dire effects, this bias can have catastrophic ramifications. Misidentification in forensic investigations may result in false allegations, and in security applications, specific populations may be unjustly singled out or denied access. In order to solve this problem, more diverse datasets must be used, and algorithms must be developed with fairness standards in place. These problems can be lessened by routinely

reviewing biometric systems for bias and including a variety of demographic groups in the testing process.

– **Legal Considerations:** Particularly in forensic and government use cases, fingerprint biometric technologies have a wide range of legal ramifications. The reliability and admissibility of biometric evidence in court is one of the main issues. Strict legal requirements must be met by forensic fingerprint identification systems in order for the evidence they produce to be trustworthy and supported by science. But many contemporary biometric algorithms are "black-box," especially those that rely on deep learning, which makes it challenging to understand how they arrive at a given judgement. This raises concerns about the accountability and transparency of these systems. Furthermore, there are jurisdictional differences in the legality of biometric data collecting, thus it is crucial to make sure local regulations are followed. While some nations have more permissive rules, others have tight legislation requiring express authorization for the use of biometric data. The ethical application of biometric technology must be covered by legal frameworks, with an emphasis on data protection, user permission, and making sure that systems are operated in a way that respects people's rights.

This review study on fingerprint biometrics has some major limitations, one of them being its scant discussion of recent developments in fingerprint sensing technology and their potential applications. Although the study thoroughly discusses the current state of the art in fingerprint recognition, it does not go into great detail into cutting edge advances like new capture methods, sophisticated sensing materials, or the effects of recent advancements in fingerprint sensor downsizing. The absence of these developments implies that the evaluation may not accurately reflect the state of the field today and in the future, given the potential impact they could have on the precision, usability, and integration of fingerprint biometrics in a variety of scenarios. This gap may allow significant technical trends to go unnoticed, along with the implications they have for research and real-world applications.

8. Conclusion

The adoption of creative solutions to current problems is essential to the future viability of biometric authentication. This entails creating reliable, scalable algorithms that can efficiently handle a variety of biometric data. It is imperative to establish uniform standards and evaluation procedures to guarantee impartial assessments of diverse authentication techniques. Strict ethical guidelines and openness campaigns are necessary to solve issues with algorithmic bias and data privacy in order to win over the public's trust. Multimodal fusion techniques that integrate various biometric features can greatly improve dependability. Furthermore, maximizing hardware technological developments is critical to increasing accuracy and productivity. The goal of research should be to develop biometric systems that are easy to use and prioritize both regulatory compliance and privacy protection. Collaboration between academic researchers, business personnel, and regulatory organizations is essential to spur innovation and guarantee the responsible application of biometric technologies across various sectors. By implementing these tactics, biometric authentication will be able to advance, satisfy the increasing need for reliable and secure authentication methods, and improve security and user experience across a range of applications.

Conflict of Interest The authors declare no conflict of interest.

References

[1] R. Ramotowski, *Lee and Gaensslen's advances in fingerprint technology*, CRC press, 2012.

[2] Y. Xu, G. Lu, Y. Lu, D. Zhang, "High resolution fingerprint recognition using pore and edge descriptors", *Pattern Recognition Letters*, vol. 125, pp. 773–779, 2019, doi:10.1016/j.patrec.2019.08.006.

[3] P. Tertychnyi, C. Ozcinar, G. Anbarjafari, "Low-quality fingerprint classification using deep neural network", *IET Biometrics*, vol. 7, no. 6, pp. 550–556, 2018, doi:10.1049/iet-bmt.2018.5074.

[4] A. A. Paulino, J. Feng, A. K. Jain, "Latent fingerprint matching using descriptor-based hough transform", *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 31–45, 2012, doi:10.1109/ijcb.2011.6117483.

[5] S. S. Arora, E. Liu, K. Cao, A. K. Jain, "Latent fingerprint matching: performance gain via feedback from exemplar prints", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2452–2465, 2014, doi:10.1109/tpami.2014.2330609.

[6] M. A. Medina-Pérez, A. M. Moreno, M. Á. F. Ballester, M. García-Borroto, O. Loyola-González, L. Altamirano-Robles, "Latent fingerprint identification using deformable minutiae clustering", *Neurocomputing*, vol. 175, pp. 851–865, 2016, doi:10.1016/j.neucom.2015.05.130.

[7] A. Sankaran, A. Jain, T. Vashisth, M. Vatsa, R. Singh, "Adaptive latent fingerprint segmentation using feature selection and random decision forest classification", *Information Fusion*, vol. 34, pp. 1–15, 2017, doi:10.1016/j.inffus.2016.05.002.

[8] W. Bian, S. Ding, W. Jia, "Collaborative filtering model for enhancing fingerprint image", *IET Image Processing*, vol. 12, no. 1, pp. 149–157, 2018, doi:10.1049/iet-ipr.2017.0059.

[9] J. Li, J. Feng, C.-C. J. Kuo, "Deep convolutional neural network for latent fingerprint enhancement", *Signal Processing: Image Communication*, vol. 60, pp. 52–63, 2018, doi:10.1016/j.image.2017.08.010.

[10] K. Cao, A. K. Jain, "Automated latent fingerprint recognition", *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 788–800, 2018, doi:10.1109/tpami.2018.2818162.

[11] R. P. Krish, J. Fierrez, D. Ramos, F. Alonso-Fernandez, J. Bigun, "Improving automated latent fingerprint identification using extended minutia types", *Information Fusion*, vol. 50, pp. 9–19, 2019, doi:10.1016/j.inffus.2018.10.001.

[12] K. Cao, D.-L. Nguyen, C. Tymoszek, A. K. Jain, "End-to-end latent fingerprint search", *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 880–894, 2019, doi:10.1109/tifs.2019.2930487.

[13] A. J. Sanchez-Fernandez, L. F. Romero, D. Peralta, M. A. Medina-Pérez, Y. Saeys, F. Herrera, S. Tabik, "Asynchronous processing for latent fingerprint identification on heterogeneous cpu-gpu systems", *IEEE Access*, vol. 8, pp. 124236–124253, 2020, doi:10.1109/access.2020.3005476.

[14] N. D. Kalka, M. Beachler, R. A. Hicklin, "Lqmetric: a latent fingerprint quality metric for predicting afis performance and assessing the value of latent fingerprints", *Journal of Forensic Identification*, vol. 70, no. 4, pp. 443–463, 2020.

[15] X. Huang, P. Qian, M. Liu, "Latent fingerprint image enhancement based on progressive generative adversarial network", "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops", pp. 800–801, 2020, doi:10.1109/cvprw50498.2020.00408.

[16] U. U. Deshpande, V. Malemath, S. M. Patil, S. V. Chaugule, "End-to-end automated latent fingerprint identification with improved dcnn-fft enhancement", *Frontiers in Robotics and AI*, vol. 7, p. 594412, 2020, doi:10.3389/frobt.2020.594412.

[17] M. Liu, P. Qian, "Automatic segmentation and enhancement of latent fingerprints using deep nested unets", *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1709–1719, 2020, doi:10.1109/tifs.2020.3039058.

[18] S. Gu, J. Feng, J. Lu, J. Zhou, "Latent fingerprint registration via matching densely sampled points", *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1231–1244, 2020, doi:10.1109/tifs.2020.3032041.

[19] D. Agarwal, A. Bansal, "A utility of pores as level 3 features in latent fingerprint identification", *Multimedia Tools and Applications*, vol. 80, no. 15, pp. 23605–23624, 2021, doi:10.1007/s11042-020-10207-x.

[20] U. U. Deshpande, V. Malemath, S. M. Patil, S. V. Chaugule, "Latent fingerprint identification system based on a local combination of minutiae feature points", *SN Computer Science*, vol. 2, no. 3, p. 206, 2021, doi:10.1007/s42979-021-00615-7.

[21] S. M. Hilles, A. Liban, O. A. Miaikil, A. M. Altrad, Y. A. B. El-Ebiary, M. M. Hilles, J. Contreras, "Latent fingerprint enhancement and segmentation technique based on hybrid edge adaptive dtv model", "2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)", pp. 8–13, IEEE, 2021, doi:10.1109/icscee50312.2021.9498025.

[22] S. M. Hilles, A. Liban, A. M. Altrad, O. A. Miaikil, Y. A. B. El-Ebiary, J. Contreras, M. M. Hilles, "Adaptive latent fingerprint image segmentation and matching using chan-vese technique based on edtv model", "2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)", pp. 2–7, IEEE, 2021, doi:10.1109/icscee50312.2021.9497996.

[23] U. U. Deshpande, V. Malemath, S. M. Patil, S. V. Chaugule, "Automatic latent fingerprint identification system using scale and rotation invariant minutiae features", *International Journal of Information Technology*, vol. 14, no. 2, pp. 1025–1039, 2022, doi:10.1007/s41870-020-00508-7.

[24] H. İ. Öztürk, B. Selbes, Y. Artan, "Minnet: Minutia patch embedding network for automated latent fingerprint recognition", "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 1627–1635, 2022, doi:10.1109/cvprw56347.2022.00169.

[25] S. Gu, J. Feng, J. Lu, J. Zhou, "Latent fingerprint indexing: Robust representation and adaptive candidate list", *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 908–923, 2022, doi:10.1109/tifs.2022.3154296.

[26] N. D. S. Cunha, H. M. Gomes, L. V. Batista, "Residual m-net with frequency-domain loss function for latent fingerprint enhancement", "2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)", vol. 1, pp. 198–203, IEEE, 2022, doi:10.1109/sibgrapi55357.2022.9991793.

[27] E. Marasco, M. He, L. Tang, Y. Tao, "Demographic effects in latent fingerprint matching and their relation to image quality", "Proceedings of the 2022 7th International Conference on Machine Learning Technologies", pp. 170–179, 2022, doi:10.1145/3529399.3529427.

[28] A. B. V. Wyzykowski, A. K. Jain, "Synthetic latent fingerprint generator", "Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision", pp. 971–980, 2023, doi:10.1109/wacv56688.2023.00103.

[29] Y. Zhu, X. Yin, J. Hu, "Fingergan: a constrained fingerprint generation scheme for latent fingerprint enhancement", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, doi:10.1109/tpami.2023.3236876.

[30] S. A. Grosz, A. K. Jain, "Latent fingerprint recognition: Fusion of local and global embeddings", *IEEE Transactions on Information Forensics and Security*, 2023, doi:10.1109/tifs.2023.3314207.

[31] R. Jindal, S. Singla, "Latent fingerprint recognition using hybrid ant colony optimization and cuckoo search.", *Int. Arab J. Inf. Technol.*, vol. 20, no. 1, pp. 19–28, 2023, doi:10.34028/iajit/20/1/3.

[32] A. B. V. Wyzykowski, A. K. Jain, "A universal latent fingerprint enhancer using transformers", *arXiv preprint arXiv:2306.00231*, 2023, doi:10.48550/arXiv.2306.00231.

[33] R. K. Dubey, J. Goh, V. L. Thing, "Fingerprint liveness detection from single image using low-level features and shape analysis", *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 7, pp. 1461–1475, 2016, doi:10.1109/tifs.2016.2535899.

[34] Z. Xia, R. Lv, Y. Zhu, P. Ji, H. Sun, Y.-Q. Shi, "Fingerprint liveness detection using gradient-based texture features", *Signal, Image and Video Processing*, vol. 11, pp. 381–388, 2017, doi:10.1007/s11760-016-0936-z.

[35] T. Chugh, K. Cao, A. K. Jain, "Fingerprint spoof buster: Use of minutiae-centered patches", *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2190–2202, 2018, doi:10.1109/tifs.2018.2812193.

[36] C. Yuan, X. Sun, Q. J. Wu, "Difference co-occurrence matrix using bp neural network for fingerprint liveness detection", *Soft Computing*, vol. 23, no. 13, pp. 5157–5169, 2019, doi:10.1007/s00500-018-3182-1.

[37] C. Yuan, X. Chen, P. Yu, R. Meng, W. Cheng, Q. J. Wu, X. Sun, "Semi-supervised stacked autoencoder-based deep hierarchical semantic feature for real-time fingerprint liveness detection", *Journal of Real-Time Image Processing*, vol. 17, pp. 55–71, 2020, doi:10.1007/s11554-019-00928-0.

[38] J. Fei, Z. Xia, P. Yu, F. Xiao, "Adversarial attacks on fingerprint liveness detection", *EURASIP Journal on Image and Video Processing*, vol. 2020, pp. 1–11, 2020, doi:10.1186/s13640-020-0490-z.

[39] Y. Zhang, C. Gao, S. Pan, Z. Li, Y. Xu, H. Qiu, "A score-level fusion of fingerprint matching with fingerprint liveness detection", *IEEE Access*, vol. 8, pp. 183391–183400, 2020, doi:10.1109/access.2020.3027846.

[40] W. Jian, Y. Zhou, H. Liu, "Densely connected convolutional network optimized by genetic algorithm for fingerprint liveness detection", *IEEE Access*, vol. 9, pp. 2229–2243, 2020, doi:10.1109/access.2020.3047723.

[41] O. F. Onifade, P. Akinde, F. O. Isinkaye, "Circular gabor wavelet algorithm for fingerprint liveness detection", *Journal of Advanced Computer Science & Technology*, vol. 9, no. 1, pp. 1–5, 2020, doi:10.14419/jacst.v9i1.29908.

[42] R. Agarwal, A. S. Jalal, K. Arya, "A multimodal liveness detection using statistical texture features and spatial analysis", *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13621–13645, 2020, doi:10.1007/s11042-019-08313-6.

[43] Y. Zhang, S. Pan, X. Zhan, Z. Li, M. Gao, C. Gao, "Fldnet: Light dense cnn for fingerprint liveness detection", *IEEE Access*, vol. 8, pp. 84141–84152, 2020, doi:10.1109/access.2020.2990909.

[44] J. Kolberg, M. Grimmer, M. Gomez-Barrero, C. Busch, "Anomaly detection with convolutional autoencoders for fingerprint presentation attack detection", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 190–202, 2021, doi:10.1109/tbiom.2021.3050036.

[45] C. Yuan, S. Jiao, X. Sun, Q. J. Wu, "Mfffld: A multimodal-feature-fusion-based fingerprint liveness detection", *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 648–661, 2021, doi:10.1109/tcds.2021.3062624.

[46] S. Agarwal, C. R. Chowdary, V. Sourabh, "Eazy learning: An adaptive variant of ensemble learning for fingerprint liveness detection", *arXiv preprint arXiv:2103.02207*, 2021, doi:10.48550/arXiv.2103.02207.

[47] S. B. Sandouka, Y. Bazi, N. Alajlan, "Transformers and generative adversarial networks for liveness detection in multitarget fingerprint sensors", *Sensors*, vol. 21, no. 3, p. 699, 2021, doi:10.3390/s21030699.

[48] A. Verma, V. K. Gupta, S. Goel, A. K. Yadav, D. Yadav, *et al.*, "Modeling fingerprint presentation attack detection through transient liveness factor-a person specific approach.", *Traitement du Signal*, vol. 38, no. 2, 2021, doi:10.18280/ts.380206.

[49] C. Yuan, Q. Cui, X. Sun, Q. J. Wu, S. Wu, "Fingerprint liveness detection using an improved cnn with the spatial pyramid pooling structure", "Advances in Computers", vol. 120, pp. 157–193, Elsevier, 2021, doi:10.1016/bs.adcom.2020.10.002.

[50] Z. İ. Özkiper, Z. Turgut, T. Atmaca, M. A. Aydın, "Fingerprint liveness detection using deep learning", "2022 9th International Conference on Future Internet of Things and Cloud (FiCloud)", pp. 129–135, IEEE, 2022, doi:10.1109/ficloud57274.2022.00025.

[51] R. C. Contreras, L. G. Nonato, M. Boaventura, I. A. G. Boaventura, F. L. Dos Santos, R. B. Zanin, M. S. Viana, "A new multi-filter framework for texture image representation improvement using set of pattern descriptors to fingerprint liveness detection", *IEEE Access*, vol. 10, pp. 117681–117706, 2022, doi:10.1109/access.2022.3218335.

[52] A. Almehmadi, "A behavioral-based fingerprint liveness and willingness detection system", *Applied Sciences*, vol. 12, no. 22, p. 11460, 2022, doi:10.3390/app122211460.

[53] S. Agarwal, A. Rattani, C. R. Chowdary, "A-ilearn: An adaptive incremental learning model for spoof fingerprint detection", *Machine Learning with Applications*, vol. 7, p. 100210, 2022, doi:10.1016/j.mlwa.2021.100210.

[54] D. Agarwal, A. Bansal, "Fingerprint liveness detection through fusion of pores perspiration and texture features", *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4089–4098, 2022, doi:10.1016/j.jksuci.2020.10.003.

[55] K. Zhang, S. Huang, E. Liu, H. Zhao, "Lfldnet: Lightweight fingerprint liveness detection based on resnet and transformer", *Sensors*, vol. 23, no. 15, p. 6854, 2023, doi:10.3390/s23156854.

[56] A. Galli, M. Gravina, S. Marrone, D. Mattiello, C. Sansone, "Adversarial liveness detector: Leveraging adversarial perturbations in fingerprint liveness detection", *IET Biometrics*, vol. 12, no. 2, pp. 102–111, 2023, doi:10.1049/bme2.12106.

[57] M. Nishanth, H. K. MR, K. N. MG, S. Kamal, T. Rao, K. Ashwini, "Fingerprint liveness detection using deep learning", "2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA)", pp. 383–388, IEEE, 2023, doi:10.1109/ciisca59740.2023.00079.

[58] Y. Myshkovskyi, M. Nazarkevych, "Robustness of fingerprint liveness detection based on convolutional neural networks", *https://ceur-ws.org/Vol-3550/short13.pdf*, 2023.

[59] C. Li, J. Hu, "A security-enhanced alignment-free fuzzy vault-based fingerprint cryptosystem using pair-polar minutiae structures", *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 543–555, 2015, doi:10.1109/tifs.2015.2505630.

[60] T. Murakami, T. Ohki, K. Takahashi, "Optimal sequential fusion for multibiometric cryptosystems", *Information fusion*, vol. 32, pp. 93–108, 2016, doi:10.1016/j.inffus.2016.02.002.

[61] Z. Jin, A. B. J. Teoh, B.-M. Goi, Y.-H. Tay, "Biometric cryptosystems: a new biometric key binding and its implementation for fingerprint minutiae-based representation", *Pattern Recognition*, vol. 56, pp. 50–62, 2016, doi:10.1016/j.patcog.2016.02.024.

[62] S. Rajendran, M. Doraipandian, "Biometric template security triggered by two dimensional logistic sine map", *Procedia computer science*, vol. 143, pp. 794–803, 2018, doi:10.1016/j.procs.2018.10.387.

[63] A. A. Al-Saggaf, "Secure method for combining cryptography with iris biometrics.", *J. Univers. Comput. Sci.*, vol. 24, no. 4, pp. 341–356, 2018.

[64] G. Panchal, D. Samanta, S. Barman, "Biometric-based cryptography for digital content protection without any key storage", *Multimedia Tools and Applications*, vol. 78, pp. 26979–27000, 2019, doi:10.1007/s11042-017-4528-x.

[65] A. A. Khan, V. Kumar, M. Ahmad, "An elliptic curve cryptography based mutual authentication scheme for smart grid communications using biometric approach", *Journal of King Saud University-Computer and Information Sciences*, 2019, doi:10.1016/j.jksuci.2019.04.013.

[66] S. Sapkal, S. Kakarwal, R. Deshmukh, "Template security of multimodal biometric system with face and fingerprint images using fuzzy vault method", *CSI Journal of Computing*, vol. 3, no. 3, pp. 36–40, 2020, doi:10.21275/v5i3.nov162389.

[67] V. Rajasekar, J. Premalatha, K. Sathya, "Enhanced biometric recognition for secure authentication using iris preprocessing and hyperelliptic curve cryptography", *Wireless communications and mobile computing*, vol. 2020, pp. 1–15, 2020, doi:10.21203/rs.2.23196/v1.

[68] F. S. Babamir, M. Kırcı, "A multibiometric cryptosystem for user authentication in client-server networks", *Computer Networks*, vol. 181, p. 107427, 2020, doi:10.1016/j.comnet.2020.107427.

[69] R. A. Rajan, P. Kumaran, "Multi-biometric cryptosystem using graph for secure cloud authentication", *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 6437–6444, 2020, doi:10.3233/jifs-179724.

[70] R. Dwivedi, S. Dey, M. A. Sharma, A. Goel, "A fingerprint based crypto-biometric system for secure communication", *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 1495–1509, 2020, doi:10.1007/s12652-019-01437-5.

[71] P. S. Chanukya, T. Thivakaran, "Multimodal biometric cryptosystem for human authentication using fingerprint and ear", *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 659–673, 2020, doi:10.1007/s11042-019-08123-w.

[72] J. Peng, B. Yang, B. B. Gupta, A. A. Abd El-Latif, "A biometric cryptosystem scheme based on random projection and neural network", *Soft Computing*, vol. 25, pp. 7657–7670, 2021, doi:10.1007/s00500-021-05732-2.

[73] S. Barzut, M. Milosavljević, S. Adamović, M. Saračević, N. Maček, M. Gnjatović, "A novel fingerprint biometric cryptosystem based on convolutional neural networks", *Mathematics*, vol. 9, no. 7, p. 730, 2021, doi:10.3390/math9070730.

[74] W. El-Shafai, F. A. H. E. Mohamed, H. M. Elkamchouchi, M. Abd-Elnaby, A. Elshafee, "Efficient and secure cancelable biometric authentication framework based on genetic encryption algorithm", *IEEE Access*, vol. 9, pp. 77675–77692, 2021, doi:10.1109/access.2021.3082940.

[75] H. A. A. El-Hameed, N. Ramadan, W. El-Shafai, A. A. Khalaf, H. E. H. Ahmed, S. E. Elkhamy, F. E. A. El-Samie, "Cancelable biometric security system based on advanced chaotic maps", *The Visual Computer*, pp. 1–17, 2021, doi:10.1007/s00371-021-02276-2.

[76] O. Ouda, K. Nandakumar, A. Ross, "Cancelable biometrics vault: A secure key-binding biometric cryptosystem based on chaffing and winnowing", "2020 25th International Conference on Pattern Recognition (ICPR)", pp. 8735–8742, IEEE, 2021, doi:10.1109/icpr48806.2021.9412957.

[77] R. Sreemol, M. S. Kumar, A. Sreekumar, "Improvement of security in multi-biometric cryptosystem by modulus fuzzy vault algorithm", "2021 International conference on advances in computing and communications (ICACC)", pp. 1–7, IEEE, 2021, doi:10.1109/icacc-202152719.2021.9708136.

[78] L. A. Elrefaei, A. M. Al-Mohammadi, "Machine vision gait-based biometric cryptosystem using a fuzzy commitment scheme", *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 2, pp. 204–217, 2022, doi:10.1016/j.jksuci.2019.10.011.

[79] X. Chang, W. Li, A. Yan, P. W. M. Tsang, T.-C. Poon, "Asymmetric cryptosystem based on optical scanning cryptography and elliptic curve algorithm", *Scientific Reports*, vol. 12, no. 1, p. 7722, 2022, doi:10.21203/rs.3.rs-1148931/v1.

[80] A. Kuznetsov, D. Zakharov, E. Frontoni, L. Romeo, R. Rosati, "Deep learning based fuzzy extractor for generating strong keys from biometric face images", "2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T)", pp. 421–426, IEEE, 2022, doi:10.1109/picst57299.2022.10238643.

[81] N. Hamian, M. Bayat, M. R. Alaghband, Z. Hatefi, S. M. Pournaghi, "Blockchain-based user re-enrollment for biometric authentication systems", *International Journal of Electronics and Information Engineering*, vol. 14, no. 1, pp. 18–38, 2022, doi:10.6636/IJEIE.202206.

[82] S. D. Patil, R. Raut, R. H. Jhaveri, T. A. Ahanger, P. V. Dhade, A. B. Kathole, K. N. Vhatkar, *et al.*, "Robust authentication system with privacy preservation of biometrics", *Security and Communication Networks*, vol. 2022, 2022, doi:10.1155/2022/7857975.

[83] L. A. Abou Elazm, W. El-Shafai, S. Ibrahim, M. G. Egila, H. Shawkey, M. K. Elsaid, N. F. Soliman, H. N. AlEisa, F. E. Abd El-Samie, "Efficient hardware design of a secure cancellable biometric cryptosystem", *Intell. Autom. Soft Comput.*, vol. 36, no. 1, pp. 929–955, 2023, doi:10.32604/iasc.2023.031386.

[84] H. A. A. Eldawy, W. El-Shafai, E. E.-D. Hemdan, G. M. El-Banby, F. E. A. El-Samie, "A robust cancellable face and palmprint recognition system based on 3d optical chaos-dna cryptosystem", *Optical and Quantum Electronics*, vol. 55, no. 11, p. 970, 2023, doi:10.1007/s11082-023-04840-7.

[85] P. Kaur, N. Kumar, "Enhanced biometric cryptosystem using ear & iris modality based on binary robust independent elementary feature", "2023 6th International Conference on Information Systems and Computer Networks (ISCON)", pp. 1–6, IEEE, 2023, doi:10.1109/iscon57294.2023.10112197.

[86] A. Sedik, A. A. A. El-Latif, M. A. Wani, F. E. A. El-Samie, N. A.-S. Bauomy, F. G. Hashad, "Efficient multi-biometric secure-storage scheme based on deep learning and crypto-mapping techniques", *Mathematics*, vol. 11, no. 3, p. 703, 2023, doi:10.3390/math11030703.

[87] S. Nagaraju, R. Nagendra, S. Balasundaram, R. K. Kumar, "Biometric key generation and multi round aes crypto system for improved security", *Measurement: Sensors*, vol. 30, p. 100931, 2023, doi:10.1016/j.measen.2023.100931.

[88] M. Sandhya, M. V. Prasad, "Securing fingerprint templates using fused structures", *IET Biometrics*, vol. 6, no. 3, pp. 173–182, 2017, doi:10.1049/iet-bmt.2016.0008.

[89] A. Roy, N. Memon, A. Ross, "Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems", *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2013–2025, 2017, doi:10.1109/tifs.2017.2691658.

[90] J. Kim, A. B. J. Teoh, "One-factor cancellable biometrics based on indexing-first-order hashing for fingerprint authentication", "2018 24th International Conference on Pattern Recognition (ICPR)", pp. 3108–3113, IEEE, 2018, doi:10.1109/icpr.2018.8545565.

[91] X. Wang, H. Li, "One-factor cancellable palmprint recognition scheme based on oiom and minimum signature hash", *IEEE Access*, vol. 7, pp. 131338–131354, 2019, doi:10.1109/access.2019.2938019.

[92] G. S. Walia, K. Aggarwal, K. Singh, K. Singh, "Design and analysis of adaptive graph-based cancelable multi-biometrics approach", *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 1, pp. 54–66, 2020, doi:10.1109/tdsc.2020.2997558.

[93] J. R. Pinto, M. V. Correia, J. S. Cardoso, "Secure triplet loss: Achieving cancelability and non-linkability in end-to-end deep biometrics", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 180–189, 2020, doi:10.1109/tbiom.2020.3046620.

[94] H. Wang, X. Dong, Z. Jin, A. B. J. Teoh, M. Tistarelli, "Security analysis of cancellable biometrics using constrained-optimized similarity-based attack", *arXiv preprint arXiv:2006.13051*, 2020, doi:10.1109/wacvw52041.2021.00012.

[95] A. K. Trivedi, D. M. Thounaojam, S. Pal, "Non-invertible cancellable fingerprint template for fingerprint biometric", *Computers & Security*, vol. 90, p. 101690, 2020, doi:10.1016/j.cose.2019.101690.

[96] U. Sharma, P. Tomar, S. S. Ali, N. Saxena, R. S. Bhadoria, "Optimized authentication system with high security and privacy", *Electronics*, vol. 10, no. 4, p. 458, 2021, doi:10.3390/electronics10040458.

[97] H. Wang, X. Dong, Z. Jin, A. B. J. Teoh, M. Tistarelli, "Interpretable security analysis of cancellable biometrics using constrained-optimized similarity-based attack", "Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision", pp. 70–77, 2021, doi:10.1109/wacvw52041.2021.00012.

[98] N. F. Soliman, A. D. Algarni, W. El-Shafai, F. E. A. El-Samie, G. Banby, *et al.*, "An efficient gcd-based cancelable biometric algorithm for single and multiple biometrics.", *Computers, Materials & Continua*, vol. 69, no. 2, 2021, doi:10.32604/cmc.2021.016980.

[99] W. Yang, S. Wang, M. Shahzad, W. Zhou, "A cancelable biometric authentication system based on feature-adaptive random projection", *Journal of Information Security and Applications*, vol. 58, p. 102704, 2021, doi:10.1016/j.jisa.2020.102704.

[100] H. Zhang, W. Bian, B. Jie, D. Xu, J. Zhao, "A complete user authentication and key agreement scheme using cancelable biometrics and puf in multi-server environment", *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 5413–5428, 2021, doi:10.1109/tifs.2021.3128826.

[101] X. Dong, Z. Jin, L. Zhao, Z. Guo, "Biocancrypto: An ldpc coded bio-cryptosystem on fingerprint cancellable template", "2021 IEEE international joint conference on biometrics (IJCB)", pp. 1–8, IEEE, 2021, doi:10.1109/ijcb52358.2021.9484391.

[102] B. Samira, R. H. Lamia, E. B. A. Najoua, "Biometric template security using watermarking reinforcement based cancellable transformation", "2021 International Conference on Cyberworlds (CW)", pp. 270–277, IEEE, 2021, doi:10.1109/cw52790.2021.00052.

[103] T. M. Dang, T. D. Nguyen, T. Hoang, H. Kim, A. B. J. Teoh, D. Choi, "Avet: a novel transform function to improve cancellable biometrics security", *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 758–772, 2022, doi:10.1109/tifs.2022.3230212.

[104] B. A. El-Rahiem, M. Amin, A. Sedik, F. E. A. E. Samie, A. M. Iliyasu, "An efficient multi-biometric cancellable biometric scheme based on deep fusion and deep dream", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2022, doi:10.1007/s12652-021-03513-1.

[105] A. M. Ayoup, A. A. Khalaf, W. El-Shafai, F. E. A. El-Samie, F. Alraddady, S. M. S. Eldin, "Cancellable multi-biometric template generation based on arnold cat map and aliasing.", *Computers, Materials & Continua*, vol. 72, no. 2, 2022, doi:10.32604/cmc.2022.025902.

[106] J. Kim, Y. G. Jung, A. B. J. Teoh, "Multimodal biometric template protection based on a cancelable softmaxout fusion network", *Applied Sciences*, vol. 12, no. 4, p. 2023, 2022, doi:10.3390/app12042023.

[107] A. M. Ayoup, A. Khalaf, F. Alraddady, F. Abd El-Samie, W. El-Safai, S. Eldin, "Selective cancellable multi-biometric template generation scheme based on multi-exposure feature fusion", *Intell. Autom. Soft Comput.*, vol. 33, no. 1, pp. 549–565, 2022, doi:10.32604/iasc.2022.024379.

[108] A. Sedik, A. A. A. El-Latif, M. El-Affendi, H. Mostafa, "A cancelable biometric system based on deep style transfer and symmetry check for double-phase user authentication", *Symmetry*, vol. 15, no. 7, p. 1426, 2023, doi:10.3390/sym15071426.

[109] S. M. S. Eldin, A. Sedik, S. S. Alshamrani, A. M. Ayoup, "Cancellable multi-biometric feature veins template generation based on sha-3 hashing.", *Computers, Materials & Continua*, vol. 75, no. 1, 2023, doi:10.32604/cmc.2023.030789.

[110] Y. Jiang, P. Shen, L. Zeng, X. Zhu, D. Jiang, C. Chen, "Cancelable biometric schemes for euclidean metric and cosine metric", *Cybersecurity*, vol. 6, no. 1, p. 4, 2023, doi:10.1186/s42400-023-00137-0.

[111] O. S. Faragallah, E. A. Naeem, W. El-Shafai, N. Ramadan, H. E.-d. H. Ahmed, M. M. A. Elnaby, I. Elashry, S. E. El-Khamy, F. E. A. El-Samie, "Efficient chaotic-baker-map-based cancelable face recognition", *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–39, 2023, doi:10.1007/s12652-021-03398-0.

[112] S. Yamamoto, H. Inaba, "Cancellable biometric authentication system by image style transfer", "2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)", pp. 794–797, IEEE, 2023, doi:10.1109/gcce59613.2023.10315291.

[113] F. Turroni, D. Maltoni, R. Cappelli, D. Maio, "Improving fingerprint orientation extraction", *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1002–1013, 2011, doi:10.1109/tifs.2011.2150216.

[114] S. Li, A. C. Kot, "An improved scheme for full fingerprint reconstruction", *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1906–1912, 2012, doi:10.1109/tifs.2012.2212012.

[115] J. Galbally, S. Marcel, J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition", *IEEE transactions on image processing*, vol. 23, no. 2, pp. 710–724, 2013, doi:10.1109/tip.2013.2292332.

[116] H.-W. Jung, J.-H. Lee, "Noisy and incomplete fingerprint classification using local ridge distribution models", *Pattern recognition*, vol. 48, no. 2, pp. 473–484, 2015, doi:10.1016/j.patcog.2014.07.030.

[117] S. Mathur, A. Vjay, J. Shah, S. Das, A. Malla, "Methodology for partial fingerprint enrollment and authentication on mobile devices", "2016 International Conference on Biometrics (ICB)", pp. 1–8, IEEE, 2016, doi:10.1109/icb.2016.7550093.

[118] T. Chugh, S. S. Arora, A. K. Jain, N. G. Paulter, "Benchmarking fingerprint minutiae extractors", "2017 International conference of the biometrics special interest group (BIOSIG)", pp. 1–8, IEEE, 2017, doi:10.23919/biosig.2017.8053498.

[119] G. S. E. Ekladious, R. Sabourin, E. Granger, "Learning global-local distance metrics for signature-based biometric cryptosystems", *Cryptography*, vol. 1, no. 3, p. 22, 2017, doi:10.3390/cryptography1030022.

[120] R. D. Labati, A. Genovese, E. Munoz, V. Piuri, F. Scotti, "A novel pore extraction method for heterogeneous fingerprint images using convolutional neural networks", *Pattern Recognition Letters*, vol. 113, pp. 58–66, 2018, doi:10.1016/j.patrec.2017.04.001.

[121] R. Gupta, M. Khari, D. Gupta, R. G. Crespo, "Fingerprint image enhancement and reconstruction using the orientation and phase reconstruction", *Information Sciences*, vol. 530, pp. 201–218, 2020, doi:10.1016/j.ins.2020.01.031.

[122] F. Liu, Y. Zhao, G. Liu, L. Shen, "Fingerprint pore matching using deep features", *Pattern Recognition*, vol. 102, p. 107208, 2020, doi:10.1016/j.patcog.2020.107208.

[123] W. Yang, S. Wang, K. Yu, J. J. Kang, M. N. Johnstone, "Secure fingerprint authentication with homomorphic encryption", "2020 Digital Image Computing: Techniques and Applications (DICTA)", pp. 1–6, IEEE, 2020, doi:10.1109/dicta51227.2020.9363426.

[124] B. T. Ahmed, O. Y. Abdulhameed, "Fingerprint authentication using shark smell optimization algorithm", *UHD Journal of Science and Technology*, vol. 4, no. 2, pp. 28–39, 2020, doi:10.21928/uhdjst.v4n2y2020.pp28-39.

[125] T. Kim, Y. Oh, H. Kim, "Efficient privacy-preserving fingerprint-based authentication system using fully homomorphic encryption", *Security and Communication Networks*, vol. 2020, pp. 1–11, 2020, doi:10.1155/2020/4195852.

[126] M. Golec, S. S. Gill, R. Bahsoon, O. Rana, "Biosec: A biometric authentication framework for secure and private communication among edge devices in iot and industry 4.0", *IEEE Consumer Electronics Magazine*, vol. 11, no. 2, pp. 51–56, 2020, doi:10.1109/mce.2020.3038040.

[127] M. Lebcir, S. Awang, A. Benziane, "Reversible watermarking technique for fingerprint authentication based on dct", *IOP Conference Series: Materials Science and Engineering*, vol. 769, no. 1, p. 012070, 2020, doi:10.1088/1757-899x/769/1/012070.

[128] D. Valdes-Ramirez, M. A. Medina-Pérez, R. Monroy, "An ensemble of fingerprint matching algorithms based on cylinder codes and mtriplets for latent fingerprint identification", *Pattern Analysis and Applications*, vol. 24, pp. 433–444, 2021, doi:10.1007/s10044-020-00911-7.

[129] I. Joshi, R. Kothari, A. Utkarsh, V. K. Kurmi, A. Dantcheva, S. D. Roy, P. K. Kalra, "Explainable fingerprint roi segmentation using monte carlo dropout", "Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision", pp. 60–69, 2021, doi:10.1109/wacvw52041.2021.00011.

[130] P. Terhörst, A. Boller, N. Damer, F. Kirchbuchner, A. Kuijper, "Midecon: Unsupervised and accurate fingerprint and minutia quality assessment based on minutia detection confidence", "2021 IEEE International Joint Conference on Biometrics (IJCB)", pp. 1–8, IEEE, 2021, doi:10.1109/ijcb52358.2021.9484404.

[131] F. Pandey, P. Dash, D. Samanta, M. Sarma, "Asra: Automatic singular value decomposition-based robust fingerprint image alignment", *Multimedia Tools and Applications*, vol. 80, pp. 15647–15675, 2021, doi:10.1007/s11042-021-10560-5.

[132] A. Muhammed, N. C. Mhala, A. R. Pais, "A novel fingerprint template protection and fingerprint authentication scheme using visual secret sharing and super-resolution", *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10255–10284, 2021, doi:10.1007/s11042-020-10095-1.

[133] R. C. Contreras, L. G. Nonato, M. Boaventura, I. A. G. Boaventura, B. G. Coelho, M. S. Viana, "A new multi-filter framework with statistical dense sift descriptor for spoofing detection in fingerprint authentication systems", "International Conference on Artificial Intelligence and Soft Computing", pp. 442–455, Springer, 2021, doi:10.1007/978-3-030-87897-9_39.

[134] M. S. El_Tokhy, "Robust multimodal biometric authentication algorithms using fingerprint, iris and voice features fusion", *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 1, pp. 647–672, 2021, doi:10.3233/jifs-200425.

[135] A. Bedari, S. Wang, W. Yang, "A secure online fingerprint authentication system for industrial iot devices over 5g networks", *Sensors*, vol. 22, no. 19, p. 7609, 2022, doi:10.3390/s22197609.

[136] A. Siswanto, A. Efendi, E. A. Kadir, "Fingerprint authentication in smart home environment based on embedded system", "2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)", pp. 1–6, IEEE, 2022, doi:10.1109/iceccme55909.2022.9988711.

[137] N. Pradeep, J. Ravi, "An revolutionary fingerprint authentication approach using gabor filters for feature extraction and deep learning classification using convolutional neural networks", "Innovations in Electronics and Communication Engineering: Proceedings of the 9th ICIECE 2021", pp. 349–360, Springer, 2022, doi:10.1007/978-981-16-8512-5_38.

[138] A. F. Y. Althabhawee, B. K. O. C. Alwawi, "Fingerprint recognition based on collected images using deep learning technology", *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, p. 81, 2022, doi:10.11591/ijai.v11.i1.pp81-88.

[139] P. Singh, H. Samuel, F. Jaafar, D. Ameyed, "Enhancing biometric security with combinatorial and permutational multi-fingerprint authentication strategies", "2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)", pp. 1–7, IEEE, 2022, doi:10.1109/picom/cbdcom/cy55231.2022.9927942.

[140] Y. Chen, Y. Yu, L. Zhai, "{InfinityGauntlet}: Expose smartphone fingerprint authentication to brute-force attack", "32nd USENIX Security Symposium (USENIX Security 23)", pp. 2027–2041, 2023.

[141] A. Popli, S. Tandon, J. J. Engelsma, A. Namboodiri, "A unified model for fingerprint authentication and presentation attack detection", "Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment", pp. 77–99, Springer, 2023, doi:10.1109/ijcb52358.2021.9484382.

[142] H. Choi, S. Woo, H. Kim, "Blind-touch: Homomorphic encryption-based distributed neural network inference for privacy-preserving fingerprint authentication", *arXiv preprint arXiv:2312.11575*, 2023, doi:10.1609/aaai.v38i20.30200.

[143] E. Marasco, M. Albanese, V. V. R. Patibandla, A. Vurity, S. S. Sriram, "Biometric multi-factor authentication: On the usability of the fingerpin scheme", *Security and Privacy*, vol. 6, no. 1, p. e261, 2023, doi:10.1002/spy2.261.

[144] R. Deshmukh, P. Yannawar, "Avao enabled deep learning based person authentication using fingerprint", "First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)", pp. 327–346, Atlantis Press, 2023, doi:10.2991/978-94-6463-196-8_26.

[145] D. Peralta, I. Triguero, R. Sanchez-Reillo, F. Herrera, J. Benitez, "Fast fingerprint identification for large databases", *Pattern Recognition*, vol. 47, no. 2, p. 588–602, 2014, doi:10.1016/j.patcog.2013.08.002.

[146] M. Lastra, J. Carabaño, P. D. Gutiérrez, J. M. Benítez, F. Herrera, "Fast fingerprint identification using gpus", *Information Sciences*, vol. 301, pp. 195–214, 2015, doi:10.1016/j.ins.2014.12.052.

[147] Y. Su, J. Feng, J. Zhou, "Fingerprint indexing with pose constraint", *Pattern Recognition*, vol. 54, pp. 1–13, 2016, doi:10.1016/j.patcog.2016.01.006.

[148] J. Khodadoust, A. M. Khodadoust, "Fingerprint indexing based on minutiae pairs and convex core point", *Pattern Recognition*, vol. 67, pp. 110–126, 2017, doi:10.1016/j.patcog.2017.01.022.

[149] D. Song, Y. Tang, J. Feng, "Aggregating minutia-centred deep convolutional features for fingerprint indexing", *Pattern Recognition*, vol. 88, pp. 397–408, 2019, doi:10.1016/j.patcog.2018.11.018.

[150] J. J. Engelsma, K. Cao, A. K. Jain, "Learning a fixed-length fingerprint representation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 6, pp. 1981–1997, 2019, doi:10.1109/tpami.2019.2961349.

[151] U. U. Deshpande, V. Malemath, S. M. Patil, S. V. Chaugule, "Cnnai: a convolution neural network-based latent fingerprint matching using the combination of nearest neighbor arrangement indexing", *Frontiers in Robotics and AI*, vol. 7, p. 113, 2020, doi:10.3389/frobt.2020.00113.

[152] V. Anand, V. Kanhangad, "Pore-based indexing for fingerprints acquired using high-resolution sensors", *Pattern Analysis and Applications*, vol. 23, pp. 429–441, 2020, doi:10.1007/s10044-019-00805-3.

[153] H. P. Singh, P. Dimri, S. Tiwari, M. Saraswat, "Segmentation techniques through machine based learning for latent fingerprint indexing and identification", 2020, doi:10.56042/jsir.v79i3.68640.

[154] I. Pérez-Sánchez, B. Cervantes, M. A. Medina-Pérez, R. Monroy, O. Loyola-González, S. García, F. Herrera, "An indexing algorithm based on clustering of minutia cylinder codes for fast latent fingerprint identification", *IEEE Access*, vol. 9, pp. 85488–85499, 2021, doi:10.1109/access.2021.3088314.

[155] Y. Xu, Y. Lu, F. Chen, G. Lu, D. Zhang, "High resolution fingerprint retrieval based on pore indexing and graph comparison", *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 226–236, 2021, doi:10.1109/tifs.2021.3139219.

[156] J. M. S. Soares, L. Barbosa, P. A. L. Rego, R. P. Magalhães, J. A. F. de Macêdo, "Using inverted index for fingerprint search", *Journal of Information and Data Management*, vol. 12, no. 5, 2021, doi:10.5753/jidm.2021.1918.

[157] G. Arora, S. Kalra, A. Bhatia, K. Tiwari, "Palmhashnet: Palmprint hashing network for indexing large databases to boost identification", *IEEE Access*, vol. 9, pp. 145912–145928, 2021, doi:10.1109/access.2021.3123291.

[158] R. Balasundaram, G. F. Sudha, "Retrieval performance analysis of multibiometric database using optimized multidimensional spectral hashing based indexing", *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 1, pp. 110–117, 2021, doi:10.1016/j.jksuci.2018.02.003.

[159] D. Osorio-Roig, T. Schlett, C. Rathgeb, J. Tapia, C. Busch, "Exploring quality scores for workload reduction in biometric identification", "2022 International Workshop on Biometrics and Forensics (IWBF)", pp. 1–6, IEEE, 2022, doi:10.1109/iwbf55382.2022.9794533.

[160] G. Arora, A. Singh, A. Nigam, H. M. Pandey, K. Tiwari, "Fkpindexnet: An efficient learning framework for finger-knuckle-print database indexing to boost identification", *Knowledge-Based Systems*, vol. 239, p. 108028, 2022, doi:10.1016/j.knosys.2021.108028.

[161] P. Drozdowski, F. Stockhardt, C. Rathgeb, C. Busch, "Signal-level fusion for indexing and retrieval of facial biometric data", *IET Biometrics*, vol. 11, no. 2, pp. 141–156, 2022, doi:10.1049/bme2.12063.

[162] D. Osorio-Roig, L. J. Gonzalez-Soler, C. Rathgeb, C. Busch, "Privacy-preserving multi-biometric indexing based on frequent binary patterns", *arXiv preprint arXiv:2310.03091*, 2023, doi:10.1109/tifs.2024.3386310/mm1.

[163] S. A. Grosz, A. K. Jain, "Afr-net: Attention-driven fingerprint recognition network", *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023, doi:10.1109/tbiom.2023.3317303.

[164] M. Kumar, D. Kumar, "An efficient gravitational search decision forest approach for fingerprint recognition", *Kuwait Journal of Science*, vol. 50, no. 2A, 2023, doi:10.48129/kjs.20635.

[165] A. K. Jain, S. S. Arora, L. Best-Rowden, K. Cao, P. S. Sudhish, A. Bhatnagar, Y. Koda, "Giving infants an identity: Fingerprint sensing and recognition", "Proceedings of the Eighth International Conference on Information and Communication Technologies and Development", pp. 1–4, 2016, doi:10.1145/2909609.2909612.

[166] A. K. Jain, S. S. Arora, K. Cao, L. Best-Rowden, A. Bhatnagar, "Fingerprint recognition of young children", *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1501–1514, 2016, doi:10.1109/tifs.2016.2639346.

[167] J. J. Engelsma, D. Deb, K. Cao, A. Bhatnagar, P. S. Sudhish, A. K. Jain, "Infant-id: Fingerprints for global good", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3543–3559, 2021, doi:10.1109/tpami.2021.3057634.

[168] I. Widiatmika, I. N. Piarsa, A. Syafiandini, "Recognition of the baby footprint characteristics using wavelet method and k-nearest neighbor (k-nn)", *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, vol. 12, no. 1, pp. 41–52, 2021, doi:10.24843/lkjiti.2021.v12.i01.p05.

[169] H. I. Yoshinori Koda, "Fundamental study of neonate fingerprint recognition using fingerprint classification", 2022, doi:10.1109/biosig55365.2022.9897017.

[170] N. Nelufule, Y. Moolla, S. Ntshangase, A. de Kock, "Biometric recognition of infants using fingerprints", Tech. rep., EasyChair, 2023, doi:10.1109/ictas56421.2023.10082749.

[171] T. O. Odu, T. Ogunfunmi, M. O. Olaniyan, I. A. Samuel, "Multi-instance contingent fusion for the verification of infant fingerprints", "World Conference on Information Systems for Business Management", pp. 197–207, Springer, 2023, doi:10.1155/2024/7728707.

[172] M. Bahzad, L. M. Labib, M. Elhosseini, M. Badawy, "M^{2}brtpc: A novel modified multimodal biometric recognition for toddlers and pre-school children approach", *Mansoura Engineering Journal*, vol. 48, no. 5, p. 5, 2023, doi:10.58491/2735-4202.3069.

[173] M. Shabil, H. Fadewar, "Fingerprint recognition of newborns and toddlers using pre-trained model under convolution neural networks", *Journal of Data Acquisition and Processing*, vol. 38, no. 2, p. 234, 2023, doi:10.5281/zenodo.7766329.

[174] K. Rajaram, N. B. Amma, S. Selvakumar, "Convolutional neural network based children recognition system using contactless fingerprints", *International Journal of Information Technology*, vol. 15, no. 5, pp. 2695–2705, 2023, doi:10.1007/s41870-023-01306-7.

[175] M. Kazi, K. Kale, R. S. Mehsen, A. Mane, V. Humbe, Y. Rode, S. Dabhade, N. Bansod, A. Razvi, P. Deshmukh, "Face, fingerprint, and signature based multimodal biometric system using score level and decision level fusion approaches", *IETE Journal of Research*, pp. 1–20, 2023, doi:10.1080/03772063.2023.2217784.

[176] S. A. El_Rahman, A. S. Alluhaidan, "Enhanced multimodal biometric recognition systems based on deep learning and traditional methods in smart environments", *Plos one*, vol. 19, no. 2, p. e0291084, 2024, doi:10.1371/journal.pone.0291084.

[177] H. Byeon, V. Raina, M. Sandhu, M. Shabaz, I. Keshta, M. Soni, K. Matrouk, P. P. Singh, T. Lakshmi, "Artificial intelligence-enabled deep learning model for multimodal biometric fusion", *Multimedia Tools and Applications*, pp. 1–24, 2024, doi:10.1007/s11042-024-18509-0.

[178] J. Priyani, P. Nanglia, P. Singh, V. Shokeen, A. Sharma, "Hgssa-bi lstm: A secure multimodal biometric sensing using optimized bi-directional long short-term memory with self-attention", *ECS Sensors Plus*, vol. 3, no. 1, p. 011401, 2024, doi:10.1149/2754-2726/ad1b3a.

[179] J. Samatha, G. Madhavi, "Securesense: Enhancing person verification through multimodal biometrics for robust authentication", *Scalable Computing: Practice and Experience*, vol. 25, no. 2, pp. 1040–1054, 2024, doi:10.12694/scpe.v25i2.2524.

[180] D. A. Reid, M. S. Nixon, S. V. Stevenage, "Soft biometrics: Human identification using comparative descriptions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1216–1228, 2014, doi:10.1109/tpami.2013.219.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Diptadip Maiti, is currently a research scholar in the department of Computer Science & Engineering at Techno India University, West Bengal. He did his B. Tech. in Computer Science & Engineering and M. Tech in Information Technology in 2005 and 2009 respectively.

His research interests are in Image Processing, Biometric Authentication, Machine Learning and Deep Learning.



Madhuchhanda Basak, is currently a research scholar in the department of Computer Science & Engineering at Techno India University, West Bengal. She did his B. Tech. & M. Tech. in Information Technology in 2006 and 2010 respectively.

Her research interests are in Image Processing, Biometric Authentication, Machine Learning and Deep Learning.



Debashis Das, is an Associate Professor in the department of Computer Science & Engineering at Dr. Sudhir Chandra Sur Institute of Technology & Sports Complex , West Bengal. He did his B. Tech. in Information Technology and M. Tech in Computer Science & Engineering in 2009 and 2012 respectively. He did his PhD in Computer Science & Engineering in 2017.

His research interests are in Image Processing, Biometric Authentication and Steganography.

Received: 20 August 2024, Revised: 12 September 2024, Accepted: 13 September 2024, Online: 26 September 2024
DOI: https://dx.doi.org/10.55708/js0309004

HivePool: An Exploratory Visualization System for Honey Beehive Data

Tinghao Feng*, Sophie Columbia, Christopher Campell, Rahman Tashakkori
Computer Science Department, Appalachian State University, Boone, 28608, USA
E-mails: fengt@appstate.edu / columbiase@appstate.edu / campellcl@appstate.edu / tashakkorir@appstate.edu
*Corresponding author: Tinghao Feng, Computer Science Department, Appalachian State University, Boone, 28608, USA & Email: fengt@appstate.edu

ABSTRACT: Honey bee health is crucial for global ecosystems, but traditional data analysis methods often struggle to capture the complex interplay between bee behavior and environmental factors. To bridge this gap, we developed HivePool, a novel data visualization and analysis tool designed to empower beekeepers and researchers with deeper insights into these interactions. This paper explores HivePool's functionalities, focusing on its interactive visualizations and innovative time-oriented pattern recognition for event prediction. By leveraging time series visualization techniques, HivePool allows users to explore not only static relationships between environmental variables but also how these variables change dynamically leading up to specific events within the hive. The paper showcases HivePool's effectiveness through two use cases: data-driven event exploration and example-driven event prediction. Overall, HivePool equips beekeepers and researchers with a powerful set of tools, facilitating a deeper understanding of bee behavior and environmental influences, ultimately leading to improved beehive health and management strategies.

KEYWORDS: Time Series Data Analysis, Event Visualization, Honey Beehive Data

1. Introduction

The alarming decline of honey bee colonies poses a significant threat, making it critical to unravel the mysteries behind Colony Collapse Disorder (CCD) [1]. Initiatives like the Appalachian Multipurpose Apiary System (AppMAIS) [2] have played a valuable role in collecting data on honey bee health. However, this data's sheer volume and complexity often create a roadblock for researchers trying to identify meaningful patterns. Existing analysis tools, while helpful, often fall short in their ability to uncover subtle correlations hidden within these intricate, multidimensional datasets [3].

To address this challenge, we developed HivePool, a groundbreaking visualization system that empowers users to delve deeper, exploring the intricate relationships between various time series data—temperature fluctuations, hive weight changes, and more — and specific events that occur within honey bee colonies, like swarming episodes. Unlike traditional analysis methods that may struggle to reveal these subtle connections, HivePool is specifically tailored to identify patterns between these diverse variables over time and how they relate to corresponding bee behavior.

The motivation behind HivePool stems from the recognition that existing data analysis tools often fail to illuminate the hidden connections within honey bee health datasets. Previous attempts, such as the visualization tool developed in our lab, provided some initial insights but lacked the sophistication needed to comprehensively analyze the richness of multidimensional time series data [3]. Similarly, Beevis, while offering a comparison of multiple time series alongside the video, lacked the capability to define events specific to honey bee behavior and analyze how these multidimensional datasets influence or relate to bee activity [4]. HivePool builds upon these past efforts to bridge the gap and provide a more powerful platform for exploring the intricate temporal relationships within honey beehive data.

By providing a platform to define honey beehive data events and explore the intricate interplay between time series variables and bee behavior, HivePool offers researchers and beekeepers a powerful tool for gaining a deeper understanding of honey bee health and the factors that influence it. This paper delves into the functionalities of HivePool and showcases its effectiveness in uncovering hidden correlations within honey beehive data. By leveraging advanced visualization techniques and analytical tools, HivePool empowers users to extract valuable insights from these complex datasets, ultimately contributing to a more comprehensive understanding of honey bee health and paving the way for developing effective monitoring and management strategies. Through compelling case studies and feedback from domain experts, we demonstrate HivePool's ability to unlock hidden patterns within bee data, facilitating informed decision-making in both beekeeping practices and research endeavors. We conclude by highlighting the significance of HivePool in advancing our understanding of honey bee behavior and its implications for colony health, while also outlining exciting avenues for future research and development.

2. Related Work

2.1. Honey Beehive Data Analysis

Understanding honey bee behavior hinges on effective data collection and analysis from beehives. Here, we ex-

plore prior research that has significantly contributed to these efforts. The AppMAIS and Beemon projects [5, 6, 7, 8] serve as prime examples, collecting diverse data (humidity, temperature, weight, audio, and video recordings) from honey beehives in Western North Carolina over the past thirteen years. This data provides a valuable foundation for further analysis.

Researchers have increasingly utilized sensor data to analyze honey bee behavior in relation to environmental factors. In [9], the authors proposed a method using a supervised machine learning classification algorithm to estimate colony health status and predict imminent collapsing states for beekeepers. In [10], the authors introduced a honey bee colony state identification solution using temperature data and fuzzy logic. In [11], the authors developed a multisensor platform for real-time and long-term measurement of relevant hive parameters and environmental conditions. While these studies establish the groundwork for hive data collection and basic analysis, they often lack comprehensive data visualization capabilities.

Visualization techniques play a crucial role in assisting honey bee analysis by leveraging human visual perception to analyze biological features or behaviors. In [12], the authors developed an automated 3D visual system that utilized two video cameras to observe and output 3D space trajectories and inflight statistics of target honey bees. In [13], the researchers visualized honey bee activity at hive entrances to monitor hive conditions and predict honey bee behavior. However, these approaches often lack an efficient visualization system with a user-friendly interface and the capability for comprehensive visual analysis, motivating the development of HivePool.

Detection and prediction of honey bee swarms have also attracted significant research interest, further motivating HivePool's development. In [14], the authors presented a bee-condition-monitoring system incorporating a deep learning process to detect bee swarming. In [15], the authors proposed a method to detect swarming remotely. In [16], the authors disclosed a method for monitoring and predicting the swarming process within honey bee colonies using vibro-acoustic information, similar to the approach taken by Beemon.

HivePool builds upon these efforts by attempting to model the pattern using a time series of all variables preceding the swarm event. This enables prediction based on historical data and identified patterns.

2.2. Multivariate Time Series Visualization

Time series data visualization has garnered significant interest among researchers, inspiring the development of HivePool. Several existing visualization tools and techniques have influenced HivePool's design, particularly in the context of exploring multivariate time series data.

TimePool, a tool for visualizing multiple univariate time series, utilizes line charts – a concept similarly adopted by HivePool to represent time series of selected variables for multiple hives [17, 18, 19]. However, displaying numerous line charts can lead to clutter, hindering data comparison and individual item tracking [20, 21]. Variations like 3D line graphs and braided graphs offer minimal improvement and can be misleading with more data [22, 23]. ChronoLenses, which applies data transformations and displays results in line charts, can also be more cluttered than the original data [24].

To address clutter from multiple line charts, HivePool employs juxtaposed techniques, projecting them onto a set of time lines for each hive [17]. Similar approaches include line graph explorers [25], two-tone pseudo coloring [26], Spark Cloud [27], and Horizon Graphs [28], all of which display time series side-by-side to avoid overlapping data. Works like Interactive Horizon Graphs [29] allow users to adjust the baseline for pattern discovery, while Qualizon graphs [30] incorporate qualitative categories with bands. In [31], the researcher present time series with multiple detail levels and abstractions, combining quantitative and qualitative representations. However, juxtaposed techniques become less effective as the number of displayed time series increases due to cognitive overload [23, 32, 33].

Juxtaposed timelines are commonly used for event exploration in various domains like news analysis [34], communication analysis [35, 36], and medical research [37, 38, 39]. HivePool leverages juxtaposed timelines to facilitate analyses of time series data for multiple hives, applying them in a different context than existing works. Similar to those used in the matrix or line-based visualizations, sorting techniques are also employed to reveal patterns within the data [40, 41].

EVis serves as another significant inspiration for HivePool [42]. It visualizes a multivariate time series preceding an event on a RadViz space. HivePool borrows this concept to define the Event Leash on the scatter plot for two environmental variables.

In [43], the author presented a set of line charts, one for each dimension, to visually analyze multivariate time series. Selecting a series in one chart would highlight it in all charts, enabling interactive exploration. While intuitive, this approach doesn't explicitly reveal temporal relationships among the dimensions.

In [44], the authors employed a two-step dimension reduction process to project a group of multivariate time series as points in a 2D space. Users could examine feature contributions to clusters within the projection space. However, reducing each time series to a single point made it difficult to compare temporal trends in detail for a large number of series. In [45], the authors used animations and trajectories to show how values of multiple objects change over time in a shared 2D space. This approach can suffer from clutter and change blindness issues with many objects.

Specifically, several methods have been proposed to visualize a single multivariate time series. In [46], the author divided a multivariate time series into time slices and projected them onto an MDS plot based on their similarities. The temporal order was mapped to the color of the projections. Time Curves [47] projected high-dimensional temporal data points to a 2D space and connected them by temporal order with a time curve. Characteristics of the time curve were extracted to reveal patterns in the evolving data. In [48], the authors projected snapshots of a dynamic network to dots in a 2D space and connected them using a curve. However, visualizing a large number of time series was not a focus of these approaches.

The trajectory-based financial time series visualization [49] is relevant to HivePool's Event Leash concept. It plotted 2D time series as trajectories in a 2D space and clustered

them based on high-dimensional feature vectors using a Self-Organizing Map [50].

2.3. Trajectory Similarity Search

Calculating the similarities of trajectories is a common computational task in many domains [51]. For our application, trajectory similarity is key to revealing similar patterns of change in environmental variables preceding specific events within a hive. HivePool, utilizing a domain appropriate similarity metric, searches for times within beehives with similar changes selected measured variables.

In [52], the authors provide an overview of trajectory similarity measures. Popular approaches vary in complexity, susceptibility to noise, and whether they account for temporal differences [53]. Existing literature divides these measures into being either metric or nonmetric. A distance measure is considered metric if it is unique, symmetric, non-negative, and satisfies the triangle inequality [52]. Nonmetric measures do not meet these requirements.

Euclidean distance is perhaps the simplest measure of quantifying similarity between trajectories [54]. However, it requires trajectories to be of equal length, which may be too primitive for some applications. Fréchet distance and Hausdorff distance both measure shape similarity [55, 56], though unlike Hausdorff, Fréchet is a metric measure [52]. Yet, none of these three distance measures account for temporal similarity.

Selected spatio-temporal similarity measures include Dynamic Time Warping and Longest Common Subsequence. Dynamic Time Warping performs a recursive search to locate the most similar points between trajectories [52, 57]. The longest Common Subsequence is a nonmetric measure that emphasizes robustness to noise [52, 58].

Similarity searching, finding the most similar trajectories within a large subset, can prove to be a computationally complex calculation. Searching large databases with many trajectories typically leads to a prohibitive runtime [59]. One attempt to devise an efficient searching algorithm is Wang and Liu's algorithm, which supports queries personalized to users [59]. Supporting such queries provides a tailored user experience, but this algorithm does not account for temporal differences.

Moreover, unlike in many real-world scenarios, notably with GPS data where trajectories may not be of the same length [60], the same number of points between compared trajectories can be guaranteed for our data. Searches that accommodate trajectories of varying lengths are unnecessarily complex for our application. Frentzos and Theodoridis proposed a novel algorithm to efficiently conduct an index-based similarity search, which utilizes R-trees [61]. Torch is a trajectory search engine that features a trajectory similarity search and utilizes R-trees [62]. Torch supports various search queries, allowing users to investigate patterns among trajectories. Likewise, HivePool supports two versions of similarity searching to find the most similar trajectories between all variables or two user-selected variables.

3. Domain Background And Data

3.1. Data Collection Methodology

This research utilizes data collected through the App-MAIS Beemon system [63]. Beemon functions by acquiring data via a Raspberry Pi 4B (Rpi) and transmitting it to a designated server for primary analysis and visualization. Fig 1 provides a comprehensive view of the Beemon system's hardware configuration.

Beemon System Hardware:

The core components of the Beemon system include a Raspberry Pi 4B (Rpi), a Raspberry Pi Camera Model 2, and protective casing for both units.

Data Acquisition:

- *Environmental Sensors:* Data pertaining to scale readings, humidity, and temperature is collected by the Rpi through Field-Programmable Gate Array (FPGA) pins.
- *Audio Data:* A microphone connected to a USB port on the Rpi captures audio information.
- *Video Analysis:* Computer vision techniques are applied to video recordings to estimate bee activity levels (VideoFileSize) and bee traffic (departures and arrivals).

Data Used for HivePool:

Internal hive data:

- *RMSDb:* Noise level within the hive in Decibel scale
- *Temperature:* Internal hive temperature
- *Humidity:* Internal hive humidity

External hive data:

- *Hive Scale:* Weight of the hive
- *VideoFileSize:* Indicator of bee activity level

Derived data:

- *Arrivals:* How many bees come into the hive entrance within the video recording
- *Departures:* How many bees fly away from the hive entrance within the video recording
- *Difference:* The difference between Arrivals and Departures

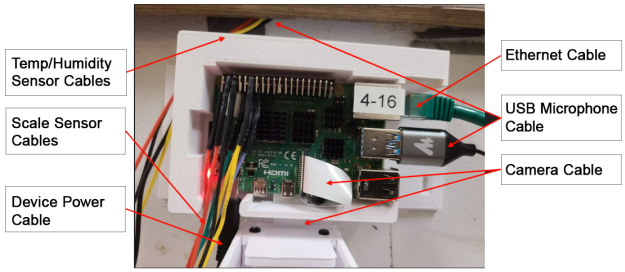


Figure 1: The Beemon system

This breakdown clarifies the data collection process and separates Beemon system details from the specific data used by HivePool for analysis.

3.2. Data Collection Timeframe

Our data collection focused on the active period of the hives, primarily daylight hours between 8:00 AM and 8:00 PM Eastern Standard Time (EST). For audio and video data, we captured one-minute segments at frequent intervals. Specifically, the Raspberry Pi (RPi) uploaded one-minute audio and video recordings every five minutes throughout the designated collection window, starting at 8:00 AM EST.

3.3. Challenges and Considerations

The Beemon system has undergone extensive testing for reliability under various weather conditions. However, occasional data gaps exist due to sensor malfunctions, power outages, or internet disruptions. Additionally, it's important to note that while video recordings capture the hive entrance, audio recordings come from within. This means they are not perfectly synchronized despite being captured simultaneously. Furthermore, propolis, a resin used by bees to seal their hives, can sometimes coat sensors and microphones, potentially affecting humidity, temperature, and audio data quality.

Despite these limitations, the data collected using the AppMAIS Beemon system provides a comprehensive snapshot of honey bee activity and environmental conditions. These datasets offer valuable insights into bee behavior and hive dynamics, even with inherent challenges. They serve as a strong foundation for developing advanced analytical and visualization techniques to unlock the complexities of honey bee ecosystems.

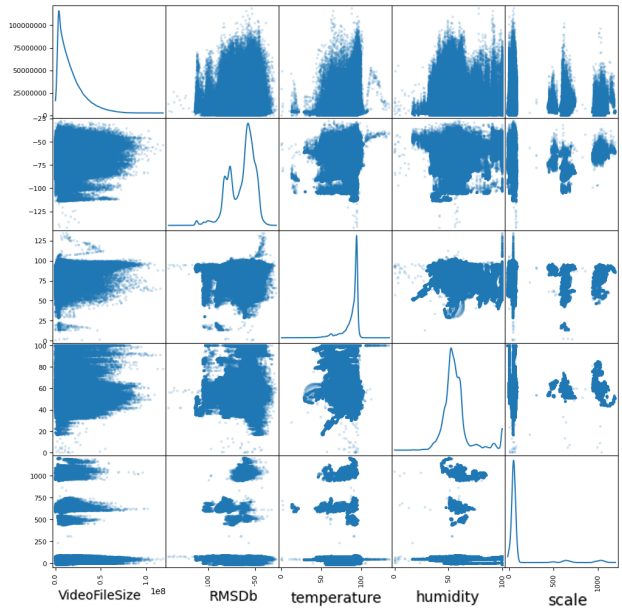


Figure 2: A scatter plot matrix with VideoFileSize, RMSDb, Temperature, Humidity, and hive Scale that shows no obvious relationships between these variables.

4. Motivation And Tasks

4.1. Motivation

Unveiling the intricate dance between honey bee behavior and environmental factors is critical for understanding colony health and dynamics. Traditional data analysis methods, however, often struggle to capture these subtle connections within complex, multidimensional datasets. Here, the rich data collected by the AppMAIS Beemon system presents a unique opportunity to delve deeper into these relationships.

While an initial scatter plot matrix (Fig 2) might not reveal clear correlations between independent variables like video file size, audio loudness (RMSDb), temperature, humidity, and hive weight, the potential to uncover hidden patterns remains significant.

This research is driven by two key motivations:

4.1.1. Understanding the Impact of Environmental Changes

We aim to discover how fluctuations in environmental variables (e.g., temperature, humidity) influence honey bee behavior within the hive. By analyzing these relationships, we can gain valuable insights into bee health and colony well-being.

4.1.2. Predicting Honey Bee Swarms

Swarming is a natural process where a portion of a colony leaves to establish a new one. Predicting swarm events is crucial for beekeepers as it allows proactive management and minimizes potential honey production losses. However, accurately predicting swarms remains challenging due to the complex interplay of factors involved.

4.2. Tasks

Addressing these motivations translates to two primary tasks for HivePool:

4.2.1. Task 1: Data-Driven Event Exploration

This task explores user-defined events within the beehive data and identifies potential correlations with environmental factors. It comprises three subtasks:

- **(T1-1) Events Definition:** Since there are no predefined behavioral events within the data, users can leverage variables like VideoFileSize, RMSDb, and bee traffic data to define events of interest. This might involve investigating variable distributions across hives and time to identify potential inflection points or thresholds that could be considered events.
- **(T1-2) Environmental Pattern Definition:** Once an event is defined, the goal is to explore the environmental factors (temperature, humidity, etc.) and how they change leading up to the event. This involves identifying patterns in these variables that might be associated with the chosen event.
- **(T1-3) Potential Event Exploration:** HivePool then quantifies the relationship between defined events and the time series of independent variables through correlation analysis. Assessing the strength and direction of these correlations helps us understand the influence of environmental factors on bee behavior and hive dynamics.

4.2.2. Task 2: Example-Driven Swarm Prediction

This task tackles the challenge of predicting honey bee swarms using historical data. It consists of the following subtasks:

- **(T2-1) Training and Test Set Splitting:** Swarm event data, typically collected by domain experts through video recordings, audio analysis, and direct observation, is the foundation for this task. The first step involves splitting these swarm samples into training and test sets. The training set defines the characteristic time series pattern preceding a swarm, while the test set is used to evaluate the prediction accuracy.
- **(T2-2) Multi-Variable Time Series Pattern Definition:** Since swarm samples are often limited, directly applying machine learning techniques might be impractical. To address this, HivePool utilizes the concept of a multi-variable time series pattern. This pattern, visualized using Detailed Leashes, depicts the changes in all relevant variables (temperature, humidity, etc.) leading up to a swarm event within the training set.
- **(T2-3) Similarity Search:** With the training pattern established, HivePool performs a similarity search across new datasets. This search identifies data points exhibiting similar patterns to the training swarm examples, potentially indicating upcoming swarms within the beehives.
- **(T2-4) Evaluation:** The effectiveness of these predictions is then evaluated using the predefined test set. By comparing the similarity search results with the known swarm events in the test set, HivePool can assess its accuracy in predicting future swarms.

This breakdown clarifies the importance of swarm prediction for beekeepers and highlights the challenges involved. It also details how HivePool addresses these challenges through its multi-variable time series pattern approach and evaluation using a test set.

5. HivePool Visualization

HivePool prioritizes user-friendliness with its comprehensive visualization design. This design allows users to gain intuitive insights into how honey bee behavior interacts with environmental variables. The toolbox offers a variety of components, including line charts for visualizing trends, time maps for spatial representation of events, hive time bar charts for activity breakdowns within the hive, scatter plots for exploring correlations, and a powerful algorithm for uncovering similar patterns within the data.

5.1. Line Charts: Unveiling Trends in Bee Activity

Line charts are HivePool's workhorse for visualizing honey bee behavior. They offer a clear view of a chosen variable's time series data for each hive. Since bee data is time-oriented and often collected intermittently (typically 8 AM to 7 PM daily), HivePool employs a segmented timeline approach. Each day is represented by a continuous segment, allowing for scaling of daily data within it.

These line charts display the variable's basic line graph for each hive, typically the event under analysis. Users can observe how the variable changes over time, following its path across each day's segment as showin in figure 3 (1). To simplify the analysis, users can select individual hives to highlight their lines, isolating and comparing specific hive data to identify unique trends and patterns.

Interaction is key: users can zoom in or out to focus on specific timeframes or adjust the scale for different levels of detail. Hovering over data points reveals specific values, fostering deeper exploration. Ultimately, HivePool's line charts empower users to analyze time series data across multiple hives. By enabling the visualization of trends, patterns, and correlations in bee behavior over time, line charts become a powerful tool for understanding hive dynamics and informing beekeeping and research decisions.

5.2. Time Maps: Decoding Bee Activity Throughout the Day

HivePool's time maps allow users to examine the daily distribution of a chosen variable more closely. They visualize this data using bubbles arranged along a vertical line within each day segment of the timeline.

Each bubble represents a specific data point, with its position on the y-axis indicating the time of day and its size reflecting the variable's value at that moment. This allows for a quick grasp of how the chosen variable fluctuates throughout the day within a hive.

Want to focus on a specific hive? No problem! Selecting a hive filters the time map, displaying only its data for focused analysis and pattern recognition across different days.

By visually depicting the variable's distribution within each day, the time map helps users understand how it changes over time. Identifying trends, peaks, and patterns in the variable's behavior provides valuable insights into hive dynamics and the influence of environmental factors.

Seamlessly integrated with the line chart, the time map allows for comparing the variable's daily distribution with its overall trend across multiple days. This lets users correlate specific time-based events with broader patterns observed in the line chart (See Fig 3 (2) for a sample: five-day line chart and time map for VideoFileSize in hive AppMAIS12R).

HivePool's time maps go beyond the line chart, offering a granular view of a variable's daily fluctuations. By providing a deeper understanding of bee activity patterns and environmental influences, the time map empowers users to make informed decisions and conduct insightful analyses.

5.3. Hive Map: Spotlighting Activity Across Hives

HivePool's Hive Map takes a unique approach to visualizing time series data. Imagine a bar for each hive, divided into colored segments that mirror the line chart's threshold (think green for above threshold, orange for below, consistent across all visualizations).

These segments clearly show how the chosen variable is distributed across different hives over time, relative to the threshold. This makes it easy to identify periods when the variable exceeds or falls below the threshold for each hive, aiding in understanding hive dynamics and environmental influences.

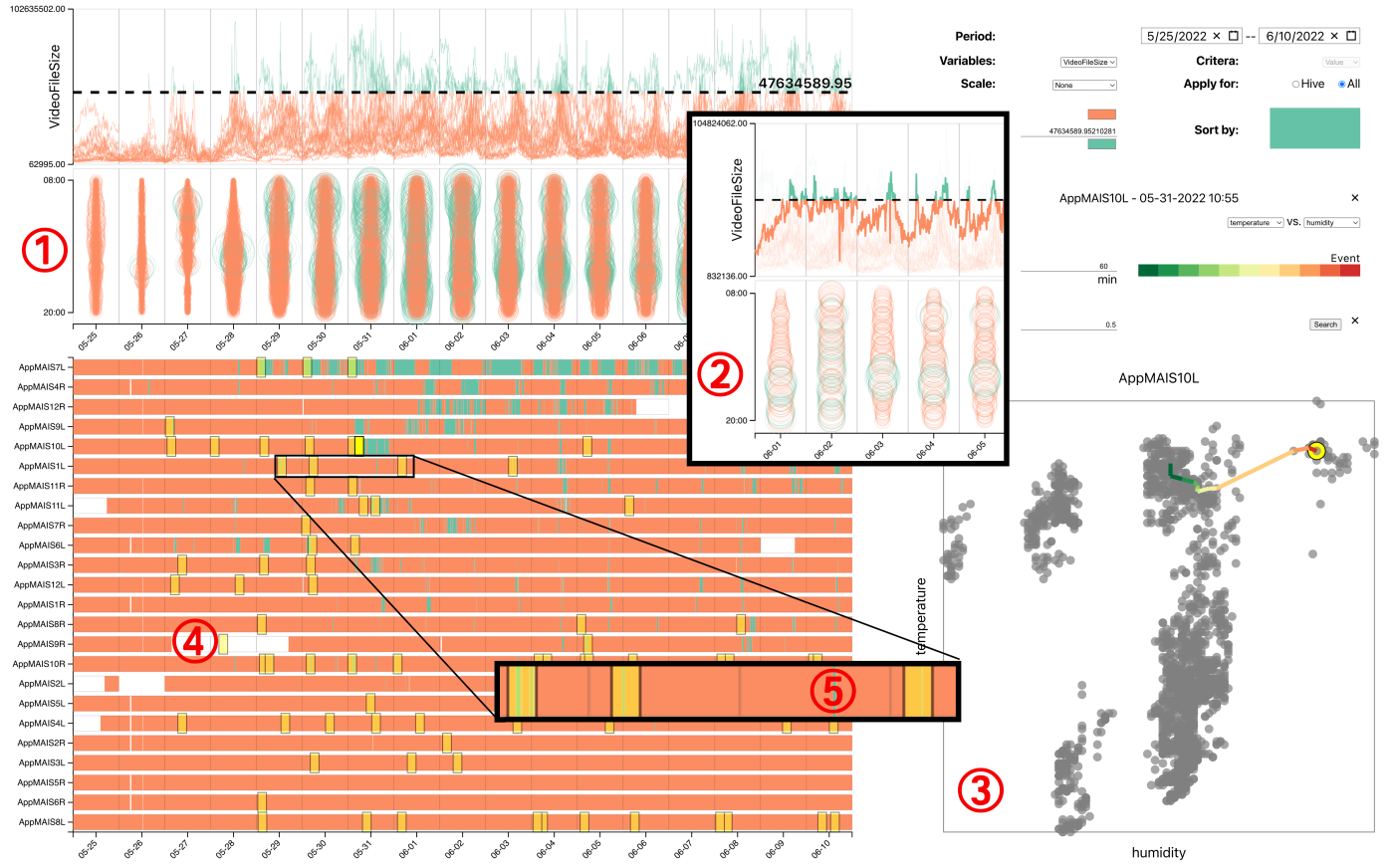


Figure 3: The layout of HivePool. (1) Line Charts and Time Maps. (2) Line Charts and Time Maps with a hive selected. (3) Scatter Plot with an Event Leash for a selected time point. (4) Hive Map. (5) Details of some data-driven event search results, which indicate the high VideoFileSize (green lines) are matched with the search results.

For better comparison, hives can be sorted based on the total length of green or orange segments, depending on your preference. This sorting helps you see which hives experience the most or least time at a specific value, revealing patterns and trends across them.

The Hive Map also boasts interactive features to enhance user experience:

- **Hive Selection:** Clicking a hive's name highlights it in the line chart and filters the time map, allowing focused analysis of a single hive's data.
- **Event Selection:** Manually select an event and see similar events highlighted on the Hive Map. This lets you explore how hives behave similarly based on specific events or conditions, offering insights into hive dynamics and environmental factors.

By providing intuitive visuals and interactive features, the Hive Map empowers users to explore the chosen variable's distribution across hives over time. This facilitates in-depth analysis of hive dynamics and environmental influences, ultimately aiding informed decision-making and a deeper understanding of honey bee behavior. Fig 3 (4) shows the Hive Map right below the Line charts and Time Map.

5.4. Scatter Plots & Event Leash: Unveiling Environmental Triggers

While line charts, time maps, and hive maps paint a vivid picture, HivePool goes beyond static visuals. It offers scatter plots to explore relationships and distributions between two chosen environmental variables for each hive. However, scatter plots lack a time dimension. Enter the Event Leash, a novel feature that bridges this gap.

Scatter Plots: These plots allow users to visualize the correlation, distribution, and potential patterns between two selected environmental variables across different hives. This empowers users to identify potential connections between, say, temperature and humidity within each hive.

Event Leash: Unveiling Time-based Triggers: This innovative feature injects time-awareness into the scatter plot view. It displays a time series of the two environmental variables leading up to a selected event (visualized on the Hive Map). Imagine selecting a time point on the Hive Map – the Event Leash showcases the trajectory of both environmental variables right before that event.

The Event Leash's path reveals how these environmental variables changed before the event. To make interpretation easier, HivePool color-codes the time order – allowing users to see the sequence of changes clearly. (Imagine a rainbow, with earlier points closer to the event being red and transitioning to violet as you move backward in time.) Fig 3 (3) shows an Event Leash on the Scatter Plot.

By integrating the Event Leash, HivePool takes scatter plots to the next level. This innovative feature allows users to go beyond static correlations and delve into the temporal dynamics of hive behavior and environmental influences. It empowers users to understand how changes in environmental variables might precede specific events within the hive.

5.5. Detailed Leash: Diving Deeper into Pre-Event Changes

The Event Leash shows how environmental variables shift before an event. But what if you want to scrutinize each variable's individual story? Here's where the Detailed Leash steps in.

The Detailed Leash provides a zoomed-in view of the time series data for each environmental variable leading up to the selected event. This detailed view allows users to see each variable's specific patterns and trends and how they might contribute to the event's occurrence.

Imagine a vertical line marking the event's exact time. The Detailed Leash plots the time series data for each variable on either side of this line, showcasing their individual pre-event behavior. Fluctuations, trends, and anomalies become clear, allowing users to understand how each variable plays a role.

By examining the Detailed Leash, users can:

- Analyze how each variable evolves over time leading up to the event.
- Identify potential correlations and causal relationships between variables and the event.
- Discover potential predictive indicators associated with the event's occurrence.

This in-depth analysis empowers users to:

- Gain a deeper understanding of the factors influencing the event.
- Make informed decisions about beekeeping practices or research interventions.

The time-oriented pattern for swarm samples in Fig 4 are defined using the Detailed Leash.

The Detailed Leash elevates HivePool's analysis capabilities. Offering a granular view of individual variable behaviors facilitates a deeper understanding of the nuanced dynamics leading up to an event. This knowledge is crucial for informed decision-making and effective interventions in beekeeping and research.

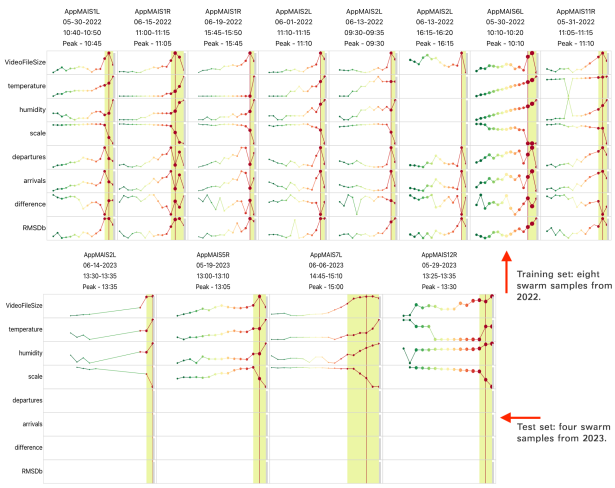


Figure 4: Detailed leashes for swarm samples. Eight samples from 2022 served as a training set and four samples from 2023 formed the test set.

5.6. Similarity Search: Unveiling Hidden Patterns in Bee Behavior

Behind the scenes, HivePool employs a greedy algorithm for similarity search. Within a user-defined date range and difference threshold, HivePool greedily searches for events with a set of leashes most similar to those of the swarm samples. An event, defined by the combination of its hive and leash start time, consists of a leash for each observed variable. By incrementally building the leashes and eliminating an event from consideration whenever its calculated difference exceeds the given threshold, HivePool avoids an expensive, exhaustive search.

5.6.1. Similarity Metric

HivePool utilizes a similarity metric that accurately calculates similarities amongst time series trajectories. Consider two leashes of the same length, a swarm and an event leash of a single variable, for which we need to calculate their trajectory similarity. HivePool calculates a similarity score between 0 and 1, where higher scores indicate more similar trajectories. The similarity between two corresponding points at index i on a swarm (s) and an event (e) leash, a and b respectively, is calculated using the following equation:

$$similarity(i) = 1 - \frac{|a-b|}{\max(\max(e),\max(s)) - \min(\min(e),\min(s))} \tag{1}$$

To avoid unexpectedly high similarity scores, we penalized differences at the head of a leash higher than those at the tail of a leash by applying weights to the calculated similarity score at each index. Let D' be an array of length n, the length of both the swarm and event leashes, where the value at each index i is given by $(i+1)^2$. The array of weights, D, is then calculated using the following formula:

$$D(i) = \frac{D'[i]}{\sum(D')} \tag{2}$$

The final similarity score, a value between 0 and 1, for two entire leashes is calculated using the following formula:

$$\sum_{i=0}^{n-1} similarity(i) * D(i) \tag{3}$$

5.6.2. Greedy Algorithm

Since there could be an excessively large number of leashes in the similarity search, HivePool pursues a greedy approach where events with difference scores that exceed the user-defined difference threshold are immediately discarded. Then, at the termination of the search, HivePool identifies a list of events whose final similarity scores are guaranteed to be below the difference threshold.

In the first step, we consider only the first point, the head, of each variable leash for an event. We compare these to the first point of each corresponding swarm variable leash to calculate an initial similarity score. If the average difference between all swarms' leashes and the event leash is above the difference threshold, we eliminate the event from further consideration for each variable.

Once this initial search concludes, we may be left with numerous overlapping leashes, or leashes that contain the same time series data from the same hive. We remove overlapping events, greedily retaining only the event with the highest initial similarity score.

We repeat the similarity calculation for each point remaining in the leashes of the events still under consideration. Similarly, if any variable's average difference exceeds the difference threshold at any calculation step, we discard that event. Once the search concludes, the events remaining are returned with each variable's calculated final similarity score.

6. Analysis Cases

This section explores HivePool's effectiveness in achieving its two core functionalities: data-driven event exploration and example-driven swarm prediction.

6.1. Case 1: Data-Driven Event Exploration

Jerry, a curious and meticulous beekeeper who is interested in understanding how internal hive conditions, specifically temperature and humidity, might influence the bees' activity levels. To delve deeper into this question, Jerry decided to leverage the power of HivePool to unlock insights from beehive data. With HivePool at his disposal, Jerry embarked on a multi-step analysis process to explore the relationship between these environmental factors and bee activity.

6.1.1. Exploring Activity and Setting the Stage

Having identified temperature and humidity as key environmental factors of interest, Jerry turned his attention to selecting a suitable indicator for bee activity within HivePool. Because a larger file size in the video recordings captured by the beehive sensors might correlate with a higher number of bees present, Jerry opted for VideoFileSize as his metric. Within HivePool, Jerry explored the VideoFileSize data using two key visualization elements: line charts and time maps. The line charts provided a detailed view of how VideoFileSize fluctuated over time for each beehive, allowing him to identify trends and patterns in bee activity across different days. The time maps, on the other hand, offered a broader overview, displaying the distribution of VideoFileSize values across the entire dataset. This visual representation enabled Jerry to quickly identify days with unusually high or low levels of bee activity.

By strategically adjusting a threshold line overlaid on the VideoFileSize line charts, Jerry could highlight specific time periods where the video file size exceeded a certain value, potentially indicating increased bee activity within the hive. This selection was then reflected on the Hive Map (Fig 3 (1) & (4)). Here, green segments represented these periods of high VideoFileSize activity for each individual hive. The hives were further sorted based on their total duration in these green segments, providing Jerry with a quick comparison of activity levels across his beehives.

However, a closer inspection of the Hive Map revealed a potential data collection bias. One particular hive (AppMAIS7L) displayed a disproportionately long green segment compared to the others. This anomaly prompted Jerry to shift his focus to another hive (AppMAIS10L) that also exhibited a substantial period of high activity (bright yellow rectangle in Fig 3 (4)). He designated this specific time period, identified as T1-1, as the event of interest for further investigation within HivePool.

6.1.2. Event Leash and Pattern Recognition

Continuing his exploration within HivePool, Jerry delved deeper into the relationship between the chosen indicator (VideoFileSize) and the environmental factors of interest (temperature and humidity). Two key visualizations offered valuable insights at this stage.

The Scatter Plot, a powerful tool for revealing correlations between variables, displayed the relationship between temperature and humidity specifically for the hive Jerry had selected (Fig 3 (3)). Each data point within the scatter plot represented a specific point in time, with its position determined by the corresponding temperature and humidity values. By visually examining the distribution of these data points, Jerry could identify potential patterns or trends in how these two environmental factors interacted.

Overlaid on top of the Scatter Plot was the Event Leash, a unique visualization element offered by HivePool (Fig 3 (3)). Represented by a rainbow line, the Event Leash served as a dynamic timeline, tracing the changes in both temperature and humidity leading up to the selected event of interest (T1-1) – the period of high VideoFileSize identified earlier. By simultaneously observing the scatter plot and the Event Leash, Jerry could gain a more comprehensive understanding of how temperature and humidity fluctuated in relation to each other and how they evolved in the lead-up to the identified period of high bee activity. In this specific instance, Jerry noticed a potentially interesting pattern – a decrease in temperature followed by an eventual rise, occurring alongside a continuous increase in humidity. This pattern, displayed prominently by the Event Leash, offered a valuable clue for Jerry to consider as he continued his investigation into the factors influencing bee activity within his hives.

6.1.3. Similarity Search and Potential Events

With the high activity period (T1-1) identified for his chosen hive (AppMAIS10L), Jerry wasn't finished yet. HivePool offered an additional layer of analysis – the ability to search for similar environmental patterns across other hives. In the background, upon selecting the event, HivePool initiated a sophisticated similarity search. This search algorithm scanned the data from all the beehives, searching for data points where the Event Leashes exhibited patterns that closely resembled the one associated with T1-1. These similar patterns could indicate similar environmental changes leading to high bee activity in other hives.

The results of this search were then displayed on the Hive Map (Fig 3 (4)) using dimmed yellow rectangles. Each rectangle represented a potential high activity event in another hive, identified by the similarity search algorithm based on the environmental patterns observed in T1-1. This visual representation on the Hive Map allowed Jerry to see how the environmental changes (represented by the green

segments associated with high VideoFileSize in his initial event) aligned with the potential high activity events in other hives (represented by the dimmed yellow rectangles) (Fig 3 (5)). The spatial distribution of these potential events offered valuable insights into the potential influence of these environmental changes across his apiary.

Furthermore, HivePool's interactive nature allowed Jerry to delve deeper. By clicking on each potential event highlighted on the Hive Map, he could access the corresponding Event Leash for that specific data point. Examining these Event Leashes (Figure 5) provided a detailed view of how temperature and humidity evolved in those other hives leading up to their potential periods of high activity. Interestingly, all four potential event Leashes displayed a similar pattern to the one Jerry observed in T1-1: a drop in temperature followed by a rise, alongside a continuous increase in humidity. This remarkable consistency across multiple hives strengthened the potential connection between this specific environmental pattern and periods of high bee activity.

This exploration gave Jerry valuable insights into the potential relationship between temperature, humidity, and bee activity within his hives. HivePool's interactive visualizations and powerful similarity search functionality empowered him to analyze user-defined events and identify potential correlations with environmental factors across his entire apiary. This comprehensive approach offered by HivePool can be a valuable tool for beekeepers seeking to understand and manage the complex interplay between environmental conditions and bee behavior.

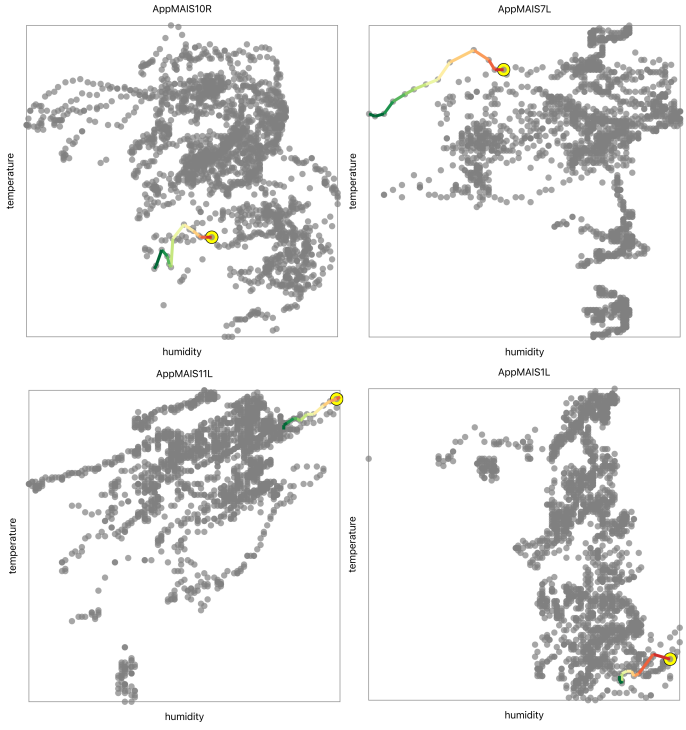


Figure 5: Four search results with high VideoFileSize for the Data-Driven Event Exploration. All the leashes show a similar pattern of continuous increasing humidity and a drop-first-and-rise-then temperature.

6.2. Case 2: Example-Driven Swarm Prediction

Gigi, a seasoned honey bee researcher, has meticulously documented honey bee swarm events for years. She has identified some swarm occurrences across 2022 and 2023 through video recordings and on-site observations. Driven by a desire to manage beehive health and prevent disruptions caused by swarming proactively, Gigi embarked on a mission to develop a model capable of predicting future swarm events.

Her initial approach involved constructing a machine learning model. With this model in mind, Gigi meticulously loaded data specific to the identified swarm occurrences. The data encompassed precise timeframes relevant to each swarm event: May 30th, 2022, to June 20th, 2022, for eight samples identified in 2022, and May 15th, 2023, to June 15th, 2023, for four samples observed in 2023. Following these preparations, Gigi attempted to train the machine learning model using the 2022 dataset, evaluating its ability to predict swarms using the 2023 dataset.

However, a significant hurdle emerged during this process: extreme class imbalance within the labels. The positive labels signifying swarm samples constituted a minuscule fraction of the data. In the training set, only eight data points out of a staggering 75,628 (from 24 hives) were labeled as swarm events. Similarly, the test set has just four positive labels out of 147,351 data points (from 29 hives). This unbalanced distribution posed a substantial challenge for the machine learning algorithm. Due to the overwhelming presence of negative labels (representing non-swarm periods), the algorithm struggled to learn the intricacies of the positive class (swarm events) effectively. Recognizing this limitation and the potential for misleading results, Gigi explored alternative swarm prediction approaches. This led her to HivePool, a visualization-based analysis tool that facilitates swarm prediction using a visually-built precedent time series pattern. By leveraging HivePool's ability to identify similar environmental patterns preceding swarm events, Gigi hoped to gain valuable insights and improve her ability to predict future swarming occurrences.

6.2.1. Setting Up the Prediction Task

To initiate the analysis, Gigi first loaded the training set containing the swarm samples from 2022. HivePool responded by generating Detailed Leashes, one for each swarm sample (T2-1). These Detailed Leashes, serving as the foundation for predicting future swarms, visually represent the time series data for each swarm event. Unlike Event Leashes used in exploratory analysis, Detailed Leashes encompass all available environmental variables. This means each Detailed Leash displays a collection of time series line charts, one for every variable measured by the beehive sensors. By examining these charts, Gigi could gain a comprehensive understanding of how all environmental factors – temperature, humidity, video file size (indicating bee activity), scale (weight of the hive), honey bee traffic (if available), and RMSDb (a measure of hive vibration) – changed in the lead-up to each swarm event in the training data (Fig 4). However, a snag emerged when Gigi loaded the test set containing the 2023 swarm samples. Due to a data collection adjustment during that period, the 2023 samples lacked information on honey bee traffic and RMSDb. Fortunately, HivePool's design allows it to adapt. Despite missing some data, HivePool could still harness the remaining variables (VideoFileSize, Temperature, Humidity, and

Scale) to identify patterns within the data. These Detailed Leashes, encompassing all available variables for both the training and test sets, collectively define the multi-variable time series patterns that HivePool will utilize to predict potential swarm events (T2-2).

6.2.2. Performing the Similarity Search

Having loaded the training set, she initiated the search function. This prompted HivePool to embark on a similarity search across the 2023 data, utilizing the Detailed Leashes constructed from the eight training swarm samples from 2022 (T2-3). The search algorithm itself is designed to be adaptable. Recognizing that the 2023 data might have missing variables due to potential adjustments in data collection, the algorithm cleverly disregards missing entries. Instead, it focuses solely on the available variables that remained consistent across both datasets: VideoFileSize, Temperature, Humidity, and Scale. By comparing these four variables across the 2023 data points, the algorithm searched for instances that exhibited a high average similarity to the patterns observed in all eight training swarm Detailed Leashes. A user-defined threshold determined the level of similarity required for a data point to be considered a potential match. The outcome of this search yielded data points within the 2023 data that displayed a high degree of similarity to the environmental patterns that preceded swarm events in the training data from 2022. These identified data points, representing potential swarm events based on similar environmental changes, were then transformed into Detailed Leashes for Gigi's review. Additionally, these potential swarm events were highlighted on the Hive Map with dimmed yellow rectangles, providing a visual representation of the distribution of the hive and time. This combined visualization approach – Detailed Leashes and Hive Map markings – offered Gigi valuable insights into potential swarm occurrences within the 2023 data.

6.2.3. Evaluating the Results

Comparing these results to the actual 2023 swarm events, Gigi observed success. HivePool accurately predicted three out of four swarms (Fig 6). However, it also generated many false positives (Type I errors). 477 results points from the 147,351-point dataset were flagged, indicating potential algorithm refinement or threshold might be needed to better balance between accuracy and highlighting potential events. Despite this, Gigi was encouraged by HivePool's ability to identify actual swarms based on environmental patterns. This initial success suggests promise for beekeepers seeking to proactively manage hives and prevent disruptions caused by swarming (T2-4).

6.2.4. Future Refinement

While the high false positive rate necessitates further development, Gigi is encouraged by the three successful predictions. This case study highlights HivePool's ability to leverage historical data and time-oriented patterns to support proactive swarm management in beehives. By addressing the false positive issue through continued research, HivePool can evolve into a valuable tool for beekeepers to predict and manage honey bee swarms more effectively.

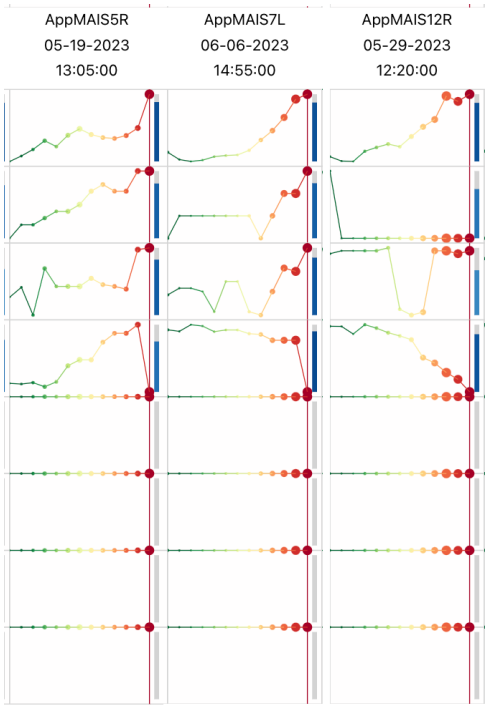


Figure 6: Swarm prediction based on data of 2023 (with missing variables) comparing to the training set successfully predict three out of four swarm samples in the test set, referring to Fig 4.

7. User Study

A user study assessed HivePool's effectiveness for beekeepers with limited data visualization experience. Participants, well-versed in honey beehive data, explored data exploration, environmental impact on bee behavior, and swarm prediction. This evaluated HivePool's accessibility and potential to empower beekeepers with data-driven insights for informed decisions, without requiring extensive training in specialized visualization techniques.

7.1. Study Design

The user study lasted approximately 30 minutes. The initial 20 minutes focused on a guided tour led by the developer to introduce participants to HivePool's core functionalities, including exploring beehive data, defining and finding potential events indicative of bee behavior, and even exploring its potential for swarm prediction. Following the guided tour, participants were empowered to take the reins for the remaining 10 minutes. This hands-on session allowed them to explore the system's features independently at their own pace. Participants were encouraged to ask questions throughout the study to ensure a productive learning environment. Finally, to gather valuable feedback on their experience, participants completed a five-question Google Form after wrapping up their exploration of HivePool.

7.2. Participants

A total of 10 participants were enrolled in the study. The group comprised 8 graduate students with a strong

background in honey beehive data and 2 experienced professors who brought their domain expertise to the table. A key characteristic shared by all participants was their lack of prior exposure to visual analytics tools. This selection process allowed us to simulate real-world scenarios where beekeepers, familiar with the intricacies of beehive data but new to visual analysis techniques, would be encountering HivePool for the first time. The user study allowed these participants to evaluate how HivePool's visual analytics approach could assist them in analyzing honey beehive data and extracting valuable insights.

7.3. User Feedback

The user study yielded valuable feedback through a Google Form survey, with questions focusing on:

- **(Q1) Did you find the system intuitive to understand?** (System Intuitiveness): This question gauged how easily participants understood HivePool's functionalities. Feedback was positive, with eight participants offering positive and one providing a neutral response. Here are some key quotes from participants: "*Surprisingly intuitive for the number of different visualizations present.*" "*I appreciated the connection between the graphs and how responsive the design was!*" "*The interface controls were very intuitive.*" "*Once you understand the layout, it is extremely informative.*""*All the labels, graphs, and charts were clear and concise.*" There is one suggestion: "*While this tool is going to be oriented towards professionals in the field, it might be helpful to add some more titles to provide quick context.*"

- **(Q2) Did you find the user interface of the visualization system easy to navigate?** (User Interface Navigation): This question assessed the user-friendliness of HivePool's interface for navigating and interacting with the data. Feedback was overwhelmingly positive. Here are some quotes: "*The placement of user-determined options on the top right was very beneficial.*" "*It gave the user very powerful tools to manipulate and browse the data.*" "*Interactive items were labeled, which is good to see.*" There is also a suggestion: "*Adding more descriptive labels to certain sections of the tool could greatly aid in navigation and usability.*"

- **(Q3) Did the visualization system provide valuable insights into honey beehive data and swarm prediction?** (Data Insights and Swarm Prediction): This question evaluated participants' perception of HivePool's effectiveness in providing valuable insights into beehive data and aiding in swarm prediction. Feedback was mostly positive, with seven participants offering praise and two providing neutral responses. Here are positive quotes: "*Swarm prediction did very well compared to the limited data about existing swarms and the number of swarms.*" "*I could see how this would be very helpful when investigating hive behavior or predicting swarms.*" "*This system's visualization helped me have more insight into the trends of the honeybee data.*" Some suggestions for improvement: "*The error rate is a little too high right now to be helpful in identifying swarms.*" "*I would like it if the line plots in the swarm detection graphs had the area under the curve colored in. This would make it a lot easier to compare the plots at first glance.*" "*I would prefer that it includes a wider range of case studies on swarm events occurring.*"

- **(Q4) How useful did you find the visualizations in understanding the data and finding patterns?** (Visualization Utility): This question explored the perceived usefulness of the visualizations in helping participants comprehend the data and identify patterns. Feedback was very positive, with eight participants offering praise and one providing a neutral response. Here are positive quotes: "*The use of similarity metrics was great.*" "*Using leashes in general to find patterns was interesting and helpful.*""*The visualizations were easy to understand while retaining detailed information.*" "*Users are able to make comparisons and draw conclusions from any of the contrasting data points.*" There are also some suggestions: "*Some sort of outlier filter may be necessary.*" "*It might be beneficial to aggregate all of the entries highlighted in blue into a visual carousel... Incorporating a search bar feature within this carousel would further improve functionality by enabling users to filter entries based on specific dates and times.*

- **(Q5) Did the visualizations aid in making informed honey beehive data analysis decisions?** (Decision-Making Support): This question assessed how well the visualizations facilitated informed decision-making about honey beehive data analysis. Feedback was overwhelmingly positive. Here are some key quotes from participants: "(The interaction) *would be very helpful in identifying sensor malfunction.*" "*The similarity search seems to be very useful...I would look at times when the queen came out of the hive.*" "*I see incredible potential in its ability to allow easy exploration through the data.*" "*Analysis of the data becomes much more reliable and easier.*""*The visualizations are the best aid in making those decisions.*" "*The insights provided by this tool would significantly boost my confidence in those choices.*"

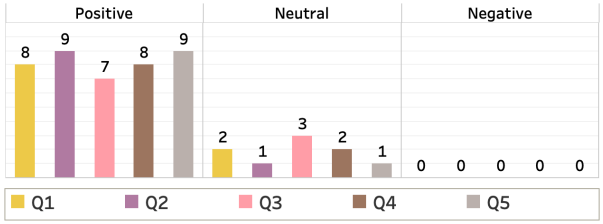


Figure 7: User feedback from the User Study.

By analyzing the survey responses and comments, we can gain valuable insights into user experience and identify areas for potential improvement within HivePool.

8. Conclusion

This work investigated the potential of HivePool, a visualization-based analysis tool, for exploring honey beehive data and predicting swarm events. We explored a case study where a beekeeper leveraged HivePool to investigate

the relationship between temperature, humidity, and bee activity within the hives. HivePool's functionalities empowered the beekeeper to visually analyze user-defined events, identify correlations between environmental factors and bee activity, and explore potential patterns across his apiary. Another case study indicates that a researcher attempted to apply machine learning for swarm prediction but encountered challenges due to class imbalance. HivePool offered an alternative approach by identifying similar environmental patterns preceding swarm events in historical data.

There is always room for improvement. Here are three key areas we will focus on in future development:

- **Incorporating Change Speed:** HivePool will incorporate "change speed" as a search criterion, allowing users to explore how rapidly environmental variables fluctuate within the hive, providing a more nuanced understanding.
- **Daily Time Series Generation:** HivePool currently uses data collected every five minutes (8am-8pm), creating time series gaps. This limits Event Leashes to the current day, hindering analysis of multi-day patterns. To address this, a daily version of HivePool will be introduced, aggregating data into daily points for continuous time series analysis of long-term environmental factors influencing bee behavior.
- **Mitigating Type I Errors:** The user study revealed high false positives in swarm prediction. To improve prediction accuracy, we'll refine the search algorithm or user thresholds, potentially incorporating additional data (honeybee traffic, RMSDb) or advanced similarity metrics.

By addressing these limitations and incorporating user feedback, HivePool will evolve into a robust, user-friendly tool that empowers beekeepers with data-driven insights into hive health, swarm prevention, and overall apiary management.

9. Acknowledgments

We gratefully acknowledge funding from the North Carolina General Assembly (NC ROI grant 2021-24) for the AppMAIS project. We also extend our appreciation to Lowe's Distinguished Professor Research Fund, relevant departments at Appalachian State University (Biology, Computer Science, Sustainability, Academic Affairs), and the undergraduate and graduate students who have contributed to the Beemon and AppMAIS data acquisition and hardware/software systems over the past eight years.

References

[1] O. US EPA, "Colony Collapse Disorder", 2013, doi:10.1007/springerreference_123456.

[2] "Appalachian Multipurpose Apiary Informatics Systems (AppMAIS)", doi:10.1007/springerreference_654321, NC ROI 2021-24.

[3] L. Richardson, C. Campell, R. Tashakkori, W. O'Brien, S. E. Davis, "Appmais simple data visualization app", "SoutheastCon 2023", pp. 366–370, 2023, doi:10.1109/SoutheastCon51012.2023.10114962.

[4] T. Feng, S. Arkle, C. Campell, R. Tashakkori, "Bee vis: A comprehensive honey bee data visualization, exploration, and analysis system", "SoutheastCon 2024", pp. 1274–1281, 2024, doi:10.1109/SoutheastCon52093.2024.10500272.

[5] R. Tashakkori, N. P. Hernandez, A. Ghadiri, A. P. Ratzloff, M. B. Crawford, "A honeybee hive monitoring system: From surveillance cameras to raspberry pis", "SoutheastCon 2017", pp. 1–7, 2017, doi:10.1109/SECON.2017.7925367.

[6] R. Tashakkori, A. S. Hamza, M. B. Crawford, "Beemon: An iot-based beehive monitoring system", *Computers and Electronics in Agriculture*, vol. 190, p. 106427, 2021, doi:10.1016/j.compag.2021.106427.

[7] D. J. Kale, R. Tashakkori, R. M. Parry, "Automated beehive surveillance using computer vision", "SoutheastCon 2015", pp. 1–3, 2015, doi:10.1109/SECON.2015.7132991.

[8] R. Tashakkori, A. Ghadiri, "Image processing for honey bee hive health monitoring", "SoutheastCon 2015", pp. 1–7, 2015, doi:10.1109/SECON.2015.7133029.

[9] A. R. Braga, D. G. Gomes, R. Rogers, E. E. Hassler, B. M. Freitas, J. A. Cazier, "A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies", *Computers and Electronics in Agriculture*, vol. 169, p. 105161, 2020, doi:10.1016/j.compag.2019.105161.

[10] A. Kviesis, V. Komasilovs, O. Komasilova, A. Zacepins, "Application of fuzzy logic for honey bee colony state detection based on temperature data", *Biosystems Engineering*, vol. 193, pp. 90–100, 2020, doi:10.1016/j.biosystemseng.2020.03.011.

[11] S. Cecchi, S. Spinsante, A. Terenzi, S. Orcioni, "A smart sensor-based measurement system for advanced bee hive monitoring", *Sensors*, vol. 20, no. 9, p. 2726, 2020, doi:10.3390/s20092726.

[12] C. Sun, P. Gaydecki, "A visual tracking system for honey bee (hymenoptera: Apidae) 3d flight trajectory reconstruction and analysis", *Journal of Insect Science*, vol. 21, no. 2, p. 17, 2021, doi:10.1093/jisesa/ieab010.

[13] T. Sledevič, V. Abromavičius, "Toward bee motion pattern identification on hive landing board", "2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)", pp. 1–4, IEEE, 2023, doi:10.1109/eStream57549.2023.10123456.

[14] G. Voudiotis, S. Kontogiannis, C. Pikridas, "Proposed smart monitoring system for the detection of bee swarming", *Inventions*, vol. 6, no. 4, p. 87, 2021, doi:10.3390/inventions6040087.

[15] A. Zacepins, A. Kviesis, E. Stalidzans, M. Liepniece, J. Meitalovs, "Remote detection of the swarming of honey bee colonies by single-point temperature monitoring", *Biosystems engineering*, vol. 148, pp. 76–80, 2016, doi:10.1016/j.biosystemseng.2016.05.013.

[16] M.-T. Ramsey, M. Bencsik, M. I. Newton, M. Reyes, M. Pioz, D. Crauser, N. S. Delso, Y. Le Conte, "The prediction of swarming in honeybee colonies using vibrational spectra", *Scientific reports*, vol. 10, no. 1, p. 9798, 2020, doi:10.1038/s41598-020-66836-2.

[17] T. Feng, Y. Hu, J. Yang, T. Polk, Y. Zhao, S. Liu, Z. Yang, "Timepool: Visually answer "which and when" questions on univariate time series", "2023 IEEE Visualization and Visual Analytics (VIS)", pp. 201–205, 2023, doi:10.1109/VIS54172.2023.00049.

[18] W. Aigner, S. Miksch, H. Schumann, C. Tominski, *Visualization of time-oriented data*, Springer Science & Business Media, 2011, doi:10.1007/978-0-85729-079-3.

[19] W. Playfair, *The commercial and political atlas: representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century*, T. Burton, 1801.

[20] B. Ondov, N. Jardine, N. Elmqvist, S. Franconeri, "Face to face: Evaluating visual comparison", *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 861–871, 2018, doi:10.1109/TVCG.2018.2865234.

[21] A. Gogolou, T. Tsandilas, T. Palpanas, A. Bezerianos, "Comparing similarity perception in time series visualizations", *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 523–533, 2018, doi:10.1109/TVCG.2018.2865235.

[22] L. Chittaro, C. Combi, G. Trapasso, "Data mining on temporal data: a visual approach and its clinical application to hemodialysis", *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 591–620, 2003, doi:10.1016/j.jvlc.2003.09.001.

[23] W. Javed, B. McDonnel, N. Elmqvist, "Graphical perception of multiple time series", *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 927–934, 2010, doi:10.1109/TVCG.2010.162.

[24] J. Zhao, F. Chevalier, E. Pietriga, R. Balakrishnan, "Exploratory analysis of time-series with chronolenses", *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2422–2431, 2011, doi:10.1109/TVCG.2011.195.

[25] R. Kincaid, H. Lam, "Line graph explorer: scalable display of line graphs using focus+ context", "Proceedings of the working conference on Advanced visual interfaces", pp. 404–411, ACM, 2006, doi:10.1145/1133265.1133331.

[26] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, T. Kaseda, "Two-tone pseudo coloring: Compact visualization for one-dimensional data", "2005 IEEE Symposium on Information Visualization", pp. 173–180, IEEE, 2005, doi:10.1109/INFVIS.2005.1532143.

[27] B. Lee, N. H. Riche, A. K. Karlson, S. Carpendale, "Sparkclouds: Visualizing trends in tag clouds", *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1182–1189, 2010, doi:10.1109/TVCG.2010.194.

[28] J. Heer, N. Kong, M. Agrawala, "Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations", "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", pp. 1303–1312, ACM, 2009, doi:10.1145/1518701.1518897.

[29] C. Perin, F. Vernier, J.-D. Fekete, "Interactive horizon graphs: improving the compact visualization of multiple time series", "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", pp. 3217–3226, ACM, 2013, doi:10.1145/2470654.2466441.

[30] P. Federico, S. Hoffmann, A. Rind, W. Aigner, S. Miksch, "Qualizon graphs: Space-efficient time-series visualization with qualitative abstractions", "Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces", pp. 273–280, ACM, 2014, doi:10.1145/2598153.2598180.

[31] R. Bade, S. Schlechtweg, S. Miksch, "Connecting time-oriented data and information to a coherent interactive visualization", "Proceedings of the SIGCHI conference on Human factors in computing systems", pp. 105–112, ACM, 2004, doi:10.1145/985692.985707.

[32] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, A. S. Wang, "Mapping color to meaning in colormap data visualizations", *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 810–819, 2018, doi:10.1109/TVCG.2018.2865147.

[33] J. Liu, A. Prouzeau, B. Ens, T. Dwyer, "Design and evaluation of interactive small multiples data visualisation in immersive spaces", "2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)", pp. 588–597, IEEE, 2020, doi:10.1109/VR46266.2020.00079.

[34] M. Krstajic, E. Bertini, D. Keim, "Cloudlines: Compact display of event episodes in multiple time-series", *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2432–2439, 2011, doi:10.1109/TVCG.2011.179.

[35] M. T. Fischer, D. Seebacher, R. Sevastjanova, D. A. Keim, M. El-Assady, "CommAID: Visual analytics for communication analysis through interactive dynamics modeling", *Computer Graphics Forum*, vol. 40, no. 3, pp. 25–36, 2021, doi:10.1111/cgf.14256.

[36] M. L. Larrea, D. K. Urribarri, "Visualization technique for comparison of time-based large data sets", M. Naiouf, E. Rucci, F. Chichizola, L. De Giusti, eds., "Cloud Computing, Big Data & Emerging Topics", pp. 179–187, Springer International Publishing, Cham, 2021, doi:10.1007/978-3-030-84825-5_16.

[37] M. Monroe, R. Lan, H. Lee, C. Plaisant, B. Shneiderman, "Temporal event sequence simplification", *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2227–2236, 2013, doi:10.1109/TVCG.2013.200.

[38] D. Klimov, Y. Shahar, M. Taieb-Maimon, "Intelligent selection and retrieval of multiple time-oriented records", *Journal of Intelligent Information Systems*, vol. 35, no. 2, pp. 261–300, 2010, doi:10.1007/s10844-008-0095-2.

[39] J. A. Fails, A. Karlson, L. Shahamat, B. Shneiderman, "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories", "2006 IEEE Symposium On Visual Analytics Science And Technology", pp. 167–174, IEEE, 2006, doi:10.1109/VAST.2006.261421.

[40] B. L. Harrison, R. Owen, R. M. Baecker, "Timelines: an interactive system for the collection and visualization of temporal data", "Graphics Interface", pp. 141–141, Canadian Information Processing Society, 1994, doi:10.20380/GI1994.18.

[41] C. J. Atman, "Design timelines: Concrete and sticky representations of design process expertise", *Design Studies*, vol. 65, pp. 125–151, 2019, doi:10.1016/j.destud.2019.10.001.

[42] T. Feng, J. Yang, M.-C. Eppes, Z. Yang, F. Moser, "Evis: Visually analyzing environmentally driven events", *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 912–921, 2022, doi:10.1109/TVCG.2021.3114867.

[43] P. Buono, A. Aris, C. Plaisant, A. Khella, B. Shneiderman, "Interactive pattern search in time series", "Visualization and Data Analysis 2005", vol. 5669, pp. 175–186, International Society for Optics and Photonics, 2005, doi:10.1117/12.587256.

[44] T. Fujiwara, N. Sakamoto, J. Nonaka, K. Yamamoto, K.-L. Ma, *et al.*, "A visual analytics framework for reviewing multivariate time-series data with dimensionality reduction", *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1601–1611, 2020, doi:10.1109/TVCG.2020.3030341.

[45] R. Takami, Y. Takama, "Visual analytics interface for time series data based on trajectory manipulation", "2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)", pp. 342–347, 2018, doi:10.1109/WI.2018.00-58.

[46] S. Cheng, K. Mueller, W. Xu, "A framework to visualize temporal behavioral relationships in streaming multivariate data", "2016 New York Scientific Data Summit (NYSDS)", pp. 1–10, 2016, doi:10.1109/NYSDS.2016.7747810.

[47] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, P. Dragicevic, "Time curves: Folding time to visualize patterns of temporal evolution in data", *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 559–568, 2016, doi:10.1109/TVCG.2015.2467851.

[48] S. van den Elzen, D. Holten, J. Blaas, J. J. van Wijk, "Reducing snapshots to points: A visual analytics approach to dynamic network exploration", *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 1–10, 2016, doi:10.1109/TVCG.2015.2468078.

[49] T. Schreck, T. Tekušová, J. Kohlhammer, D. Fellner, "Trajectory-based visual analysis of large financial time series data", *ACM SIGKDD Explorations Newsletter*, vol. 9, no. 2, pp. 30–37, 2007, doi:10.1145/1345448.1345453.

[50] T. Kohonen, *Self-organizing maps*, Springer Science & Business Media, 2012, doi:10.1007/978-3-642-56927-2.

[51] K. Toohey, M. Duckham, "Trajectory similarity measures", *Sigspatial Special*, vol. 7, no. 1, pp. 43–50, 2015, doi:10.1145/2811234.2811240.

[52] N. Magdy, M. Sakr, T. Abdelkader, K. Elbahnasy, "Review on trajectory similarity measures", 2015, doi:10.1109/IntelCIS.2015.7397286.

[53] D. Hu, L. Chen, H. Fang, Z. Fang, T. Li, Y. Gao, "Spatio-temporal trajectory similarity measures: A comprehensive survey and quantitative study", *IEEE Transactions on Knowledge and Data Engineering*, 2023, doi:10.1109/TKDE.2023.3241234.

[54] C. Faloutsos, M. Ranganathan, Y. Manolopoulos, "Fast subsequence matching in time-series databases", *SIGMOD Rec.*, vol. 23, no. 2, p. 419–429, 1994, doi:10.1145/191843.191925.

[55] H. Alt, M. Godau, "Computing the fréchet distance between two polygonal curves", *International Journal of Computational Geometry & Applications*, vol. 5, no. 01n02, pp. 75–91, 1995, doi:10.1142/S0218195995000064.

[56] M.-P. Dubuisson, A. K. Jain, "A modified hausdorff distance for object matching", "Proceedings of 12th international conference on pattern recognition", vol. 1, pp. 566–568, IEEE, 1994, doi:10.1109/ICPR.1994.576361.

[57] E. Keogh, C. A. Ratanamahatana, "Exact indexing of dynamic time warping", *Knowledge and information systems*, vol. 7, pp. 358–386, 2005, doi:10.1007/s10115-004-0154-5.

[58] M. Vlachos, G. Kollios, D. Gunopulos, "Discovering similar multidimensional trajectories", "Proceedings 18th international conference on data engineering", pp. 673–684, IEEE, 2002, doi:10.1109/ICDE.2002.994784.

[59] H. Wang, K. Liu, "User oriented trajectory similarity search", "Proceedings of the ACM SIGKDD International Workshop on Urban Computing", pp. 103–110, 2012, doi:10.1145/2346496.2346515.

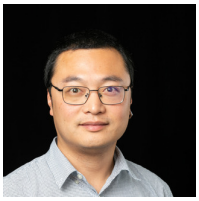
[60] W. Zheng, R. Zhou, Z. Zhang, Y. Zhong, S. Wang, Z. Wei, H. Ji, "Understanding the tourist mobility using gps: How similar are the tourists?", *Tourism management*, vol. 71, pp. 54–66, 2019, doi:10.1016/j.tourman.2018.09.011.

[61] E. Frentzos, K. Gratsias, Y. Theodoridis, "Index-based most similar trajectory search", "2007 IEEE 23rd International Conference on Data Engineering", pp. 816–825, IEEE, 2006, doi:10.1109/ICDE.2007.367926.

[62] S. Wang, Z. Bao, J. S. Culpepper, Z. Xie, Q. Liu, X. Qin, "Torch: A search engine for trajectory data", "The 41st international ACM SIGIR conference on research & development in information retrieval", pp. 535–544, 2018, doi:10.1145/3209978.3210047.

[63] R. Tashakkori, A. S. Hamza, M. B. Crawford, "Beemon: An IoT-based beehive monitoring system", *Computers and Electronics in Agriculture*, vol. 190, p. 106427, 2021, doi:10.1016/j.compag.2021.106427.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Tinghao Feng has completed his PhD in Computer Science from UNC Charlotte. He is an Assistant Professor in the Computer Science Department at Appalachian State University.

Tinghao's research primarily focuses on analyzing time-oriented data and developing visualization and analysis systems. He aims to create applications to explore temporal patterns and trends, drawing from his research experience in environmental science, healthcare, and finance. He has published numerous papers in reputable journals and conferences and has been recognized with several awards for his contributions.



Sophie Columbia completed her bachelor's degree from Appalachian State University in 2023. She completed her master's degree from Appalachian State University in 2024. Sophie worked as a Graduate Research Assistant on the AppMAIS project while she was a graduate student at Appalachian State University.



Christopher Campell has done his master's degree from Appalachian State University in 2020. He is currently a Research Associate in Machine Learning in the Computer Science Department at Appalachian State University.

Christopher Campbell leads the AppMAIS research group at Appalachian State University, focusing on innovative applications of machine learning and IoT technologies. His work includes developing advanced audio data labeling tools, estimating bee traffic using YOLO and optical flow, and creating comprehensive visualization systems for honey bee data. Chris's research also explores the health of honey beehives through cast ratio analysis at hive entrances, detecting honey bee piping with convolutional neural networks, and implementing IoT platforms for real-time beehive monitoring and alerting. Additionally, he has worked on non-invasive swarm detection methods, simple data visualization applications, and fleet management systems for apiaries using Jenkins. Through these diverse projects, Chris aims to enhance the understanding and management of honey bee populations and their environments.



Rahman Tashakkori received his bachelor's degree from Ahwaz University, Iran, in 1987. He completed two master's and PhD degrees at Louisiana State University, Baton Rouge, USA, in 1994, 1995, and 2001, respectively.

Rahman's main areas of research are in image processing, data science, honey bee monitoring systems, and computer science education. He has directed several NSF scholarship programs, including CSEMS, STEM, S-STEM< and LSAMP. He is a Senior Member of IEEE.

Received: 19 August, 2024, Revised: 20 September, 2024, Accepted: 21 September, 2024, Online: 11 October, 2024

DOI: https://doi.org/10.55708/js0310001

An Integrated Approach to Manage Imbalanced Datasets using PCA with Neural Networks

Swarup Kumar Mondal*, **Anindya Sen**

Department of Electronics and Communication Engineering, Heritage Institute of Technology, Kolkata, 700107, India

*Corresponding author: Swarup Kumar Mondal, Kolkata 700107, swarup.kumarmondal.ece24@heritageit.edu.in

ABSTRACT: Imbalanced dataset handling in real time is one of the most challenging tasks in predictive modelling. This work handles the critical issues arising in imbalanced dataset with implementation of artificial neural network and deep neural network architecture. The usual machine learning algorithms fails to achieve desired throughput with certain input circumstances due to mismatched class ratios in the sample dataset. Dealing with imbalanced dataset leads to performance degradation and interpretability issue in traditional ML architectures. For regression tasks, where the target variable is continuous, the skewed data distribution is major issue. In this study, we have investigated a detailed comparison of traditional ML algorithms and neural networks with dimensionality reduction method to overcome this problem. Principle component analysis has been used for feature selection and analysis on real time satellite-based air pollution dataset. Five regression algorithms Multilinear, Ridge, Lasso, Elastic Net and SVM regression is combined with PCA and non PCA to interpret the outcome. To address unbalanced datasets in real-time, deep neural networks and artificial neural network architectures have been developed. Each model's experiments and mathematical modelling is done independently. The Deep neural network is superior compared to other conventional models for performance measures of target variable in imbalanced datasets.

KEYWORDS: Imbalanced data, Regression, Deep Neural Network, Artificial Neural Network, Support Vector Machine

1. Introduction

In the domain of artificial intelligence (AI), the persistent challenge of imbalanced datasets remains a focal point. Accuracy, a prime indicator of model performance is primarily dependent on the balance of the data set. This issue is of paramount importance in practical applications like fraud detection [1], medical diagnosis [2], remote sensing [3], engineering [4] and anomaly detection [5], where the class of interest is typically a minority within the dataset. An uneven distribution of inputs and outputs (classes) in an imbalanced dataset is defined as one class, called the minority class, having substantially fewer instances than the other classes, called the majority class. Due to the uneven class ratios in the dataset, handling unbalanced datasets presents a challenge. In an out of balanced dataset, the majority class may make up a large portion of the training dataset, while the minority class is underrepresented in the dataset. The problem with a model trained on such out of balanced data set is that the model progressively learns to achieve high accuracy by consistently predicting the majority class, even if recognizing the minority class is equal or more important when applying the model to a real-world scenario.

Many machine learning approaches have been widely applied in a variety of applications. Machine learning models such as Multilinear Regression (MLR), Ridge Regression (RR), Lasso Regression (LR) and Elastic Net Regression (ELR) are mainly used for regression task in diverse ways. However, the traditional regression algorithms are not always beneficial to meet the requirement of large-scale raw datasets.

One of the most helpful algorithms that aims to identify underlying links in a batch of data by simulating

how the human brain functions is the neural network. Neural networks are one of the most effective methods for fitting various types of stationary and non-stationary data. A degree of balance in the dataset is essential for neural networks to function well. However, the conventional assumption of a balanced data becomes untenable for certain scenarios, where the minority class is significantly underrepresented. This imbalance leads neural networks to exhibit negative performance, particularly in accurate prediction of minority classes. The issue of data imbalance can also be addressed via deep learning synaptic models, which have been shown to considerably raise performance benchmarks in both regression and classification tasks. Neural networks topologies keep changing and getting better in 21st century, that present an opportunity to enhance the computational modeling capabilities and enabling more accurate predictions in various data-driven use cases [6].

1.1. Hypothesis

In an imbalanced stationary dataset, the neural network-based architecture works more efficiently and produce better performance accuracy over multiple regression models (MLR, RR, LR, ELR).

1.2. Objective

This research implements a novel way to apply neural networks to imbalanced real-time satellite datasets. Multiple regression model is trained for analyzing the performance for real time satellite data based on Air Quality Index (AQI) prediction. The same dataset is used for the implementation of Multilinear Regression (MLR), Ridge Regression (RR), Lasso Regression (LR), Elastic Net Regression (ELR), and Support Vector Machine Regression (SVM). In addition, the dataset is processed using Principal Component Analysis (PCA). Essentially, two neural network architectures- the Artificial Neural Network (ANN) and the Deep Neural Network (DNN) are compared with four traditional regression models and SVM. After comparison, the proposed method produced promising results which shows how to analyze the imbalanced dataset in an effective manner. Experiments are presented to demonstrate how the neural network approach efficiently improves the model accuracy and R-squared score (R2).

2. Literature Review

Earlier works which specified different approaches and methods for theoretical and practical implementation as well as handling the imbalanced dataset are discussed below:

A training procedure for Multilayer Perceptron (MLPs) is implemented using an imbalanced dataset. This article [3] uses a remote-sensing dataset that includes agricultural classes including potatoes, carrots, wheat, sugar beets, and stubble. The suggested method seeks to increase the stability of classification findings and accelerate training by an average of 41.5 times.

In order to solve Haberman's surviving unbalanced data set challenges, a modified learning method [6] for ANN is presented. The artificial neural network's output layer uses Particle Swarm Optimization (PSO) as part of its technique to optimize the step function's decision boundary. The study demonstrates increase in the average Geometric-Mean Test of classifier performance (80.16 for training and 70.47 for testing).

Three image data sets and five document data sets with different degrees of imbalance are used in an experiment [7]. It proposed a loss function called mean false error (MFE) and its enhanced counterpart, mean squared false error (MSFE), for training DNN on unbalanced data sets. Theoretical analysis is empirically confirmed, and experiments and comparisons demonstrating the superiority of the proposed approach over traditional methods in classifying imbalanced data sets on deep neural networks demonstrate the effectiveness of the proposed methods in extremely imbalanced data sets.

The creation of a system to simulate and predict maize and soybean yields on a county-by-county basis in the American Midwest and Great Plains using Artificial Neural Networks (ANN) has been studied [8]. It utilized multi-temporal remote sensing images to derive NDVI values, which characterized the entire growing process. The methodology employed a feed-forward multi-layer perceptron (MLP) neural network for learning and the SCE-UA method for training the NN. The outcomes showed that multivariate linear regression (MLR) is 20% inferior than the ANN.

The purpose of gathering seven severely unbalanced data sets [9] is to assess how well various Support Vector Machine (SVM) modelling techniques work. Various "rebalance" procedures, including cost-sensitive learning and over and under sampling, were included into SVM modelling to tackle the problem of class imbalance. It introduced GSVM-RU algorithm, which comes up as a state of art approach with 85.2 G-Mean, 91.4 AUC- ROC, 66.5 F Measure and 181 Efficiency.

The work [10] focuses on the unique problem of Deep Imbalanced Regression (DIR) and uses large-scale datasets from computer vision, natural language processing, and healthcare. Feature distribution smoothing (FDS) and label distribution smoothing (LDS) are two efficient

techniques that the study suggests using to handle unbalanced data with continuous objectives. The MAE and G-Mean improved significantly ranged from 0.1 to 2.7.

A large-scale dataset is utilized including face attribute classification and edge detection tasks, as well as controlled class imbalance in the MNIST digit classification. A novel method [11] for learning deep feature embeddings that effectively handle imbalanced data classification. The proposed approach involves quintuplet sampling and a triple-header hinge loss to enforce relationships during feature learning. The LMLE-kNN has outperformed by a large margin with a mean per class accuracy of 84 as compared to other traditional method.

To confirm the notion that models trained using Dense Loss perform better in underrepresented areas of the dataset than models built with a conventional training approach, synthetic datasets [12] with varied features, such as heavy-tailed datasets, are utilized. The findings demonstrated that MLP with Dense Loss performs better than MLP without Dense Loss, with an average Root Mean Square Value (RMSE) ranging from 1.21 to 7.02.

According to a study [13] on small and medium-sized enterprises (SMEs) in Malaysia, such as those in the transportation, storage, catering, lodging, and hotel industries, the Synthetic Minority Oversampling Technique (SMOTE) is effective in improving the classification accuracy of logistic regression models when the data are highly unbalanced. The study used a dataset of 601 failed and 26,284 failing SMEs between 1999 and 2013. The findings demonstrated that, with 57.23% sensitivity and 58.83% specificity, the SMOTE logistic regression technique produced superior metrics when compared to classical logistic regression.

SMOGN algorithm [14] combines two oversampling techniques knows as SMOTE and gaussian noise. This algorithm is used as a preprocessing solution for unbalanced regression problems with the aim of improving the performance of regression algorithms. SMOGN performs exceptionally well with multivariate adaptive regression spline (MARS) and random forest (RF) learners, demonstrating enhanced recall without appreciable loss of precision, according to tests on 20 distinct regression datasets.

3. Materials and Methods

This section introduces the imbalanced stationary data regression algorithms and the experimental data sets used in this study.

3.1. Dataset

The dataset consists of the satellite records of greenhouse gases (GHGs) emission in India. The initial dataset consists of two million five hundred eighty nine thousand and eighty three sets of reviews. It possesses a total of fifteen features as Stationid, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI and AQI_Bucket. The raw data belongs to a stationary imbalanced category. A preprocessed version of the data having five thousand and ninety two samples is utilized in this study. The dataset used in this study contains no missing features. The original dataset is open source and can be downloaded here: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?select=station_hour.csv

3.2. Operating system and software

The analytical code implementation and execution have been conducted on a 64-bit operating system with an x64-based processor system type, complemented by 8GB of RAM. Version 6.3.0 of Jupyter Notebook is utilized as an interactive programming environment. In addition, the Python script has been run on the Google Colab cloud platform, which has an Intel(R) Xeon(R) CPU operating at 2.30 GHz, 12.7 GB of RAM, and 108 GB of disc space.

3.3. Data Preprocessing

The quality of the underlying data is the primary determinant of the success of visualizations and the construction of efficient machine learning (ML) models. Preprocessing is essential for increasing data quality since it reduces noise, which boosts ML systems' processing speed and capacity for generalization. Two prevalent issues encountered in data extraction and monitoring applications are outliers and missing data. Different techniques have been employed to identify outlier values and handle not-a-number (NAN) values. Table 1 shows the visual representation of distribution of missing values across different features of the dataset. Notably, Xylene emerges as the feature with the highest number of missing values, while CO exhibits the least. To guarantee that the significance of the variables is not impacted by changes in their ranges or units, all missing values have been removed from the dataset and the standard scaler normalization process has been executed.

Table 1: Missing Values of Features.

Features	Total Missing Values
$PM_{2.5}$	647,689
PM_{10}	1,119,252
NO	553,711
NO_2	528,973
NO_x	490,808

NH_3	1,236,618
CO	499,302
SO_2	742,737
O_3	725,973
Benzene	861,579
Toluene	1,042,366
Xylene	2,075,104
AQI	570,190
AQI Bucket	570,190

3.4. Feature Selection

To select significant features, the correlation of the AQI with other features has been analyzed. Table 2 displays the precise correlation values between the AQI and each contaminant in the dataset. It is evident that the majority of the features have very little correlation. The ideal number of input variables for the creation of machine learning models has been found using a feature selection method based on Principle Component Analysis (PCA). To improve computing efficiency and accuracy, the intended objective is to project a d-dimensional dataset onto a (k)-dimensional subspace in order to minimize its dimensions.

Table 2: Correlation between AQI and Pollutants.

Sl. No.	Features	Correlation value	Sl. No	Features	Correlation value
1	$PM_{2.5}$	0.786344	7	CO	0.432508
2	PM_{10}	0.757663	8	SO_2	0.135806
3	NO	0.288469	9	O_3	0.094589
4	NO_2	0.441733	10	Benzene	0.125644
5	NO_x	0.426584	11	Toluene	0.169872
6	NH_3	0.283593	12	Xylene	0.090680

3.5. Experiment Performed

3.5.1. Proposed Architecture

Figure 1 illustrates the workflow of our proposed architecture. It begins with collecting dataset and data preprocessing. After that, it checks if there are any missing values present or not in the dataset. It there are no missing values then the non-redundant dataset go through the exploratory data analysis process. The next steps involve applying the features and target value into various ML and neural network architecture. If the result is unstable, and poor then the input goes through PCA with 2, 3 and 5 components to enhance the model stability and handle the imbalanced real time dataset. After PCA, it selects the optimal features and target values to perform data splitting and perform regression to produce the improved results. This approach enhanced the model stability, interpretability and able to overcome the issue of handling imbalanced dataset using standard regression models.

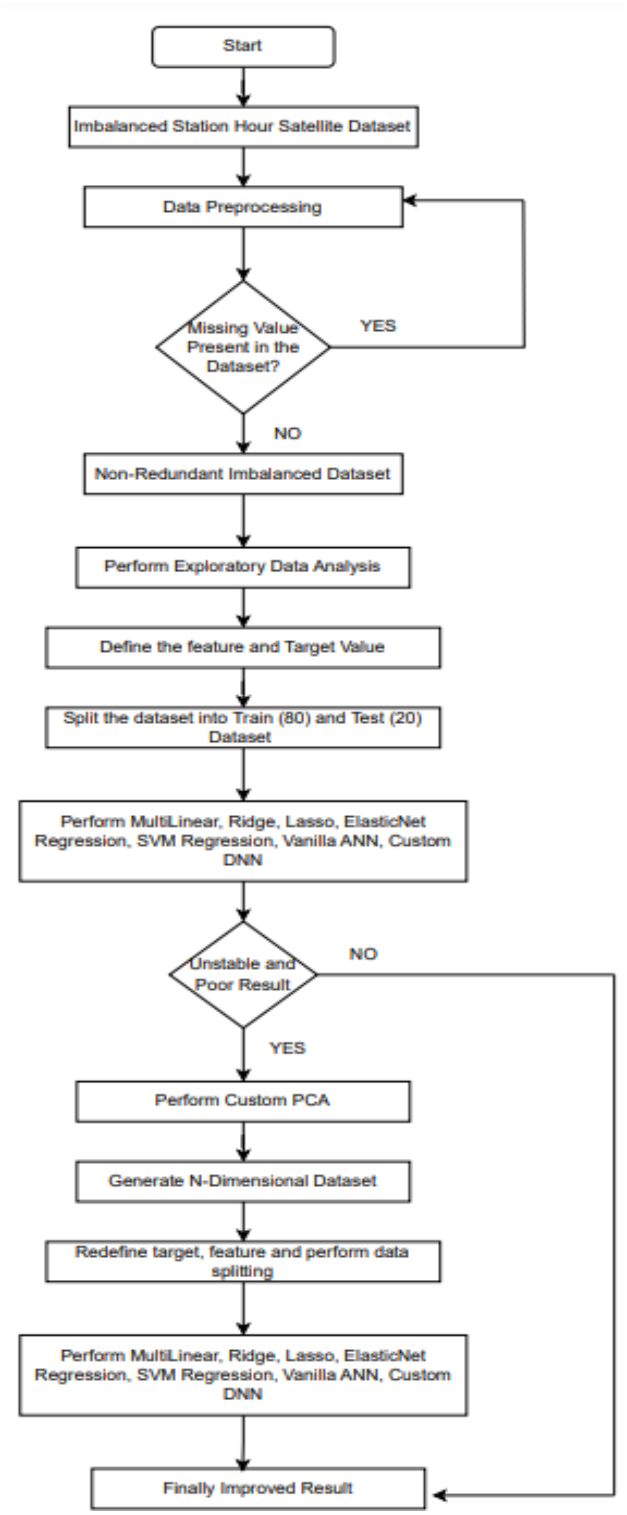


Figure 1: Process Workflow of the Proposed Architecture

3.6. Algorithm

Algorithm: PCA Combined Neural Network

Input: D → Complete dataset (Station hour Satellite dataset)

Output: Predicted AQI Value

3.7. Steps

***1)* Load and Clean Data**

Load Dataset D, handle missing values and make

the dataset non-redundant.

2) Perform Initial EDA and Model Evaluation

I. Define features and target variable.

II. Split dataset into training (80%) and testing (20%) sets.

III. Apply the following 7 models on the imbalanced dataset: MLR, RR, LR, ELR, SVM, Vanilla ANN and Custom DNN

IV. Evaluate performance metrics and check if the results are unstable or not. If the results are not good then go to step 3.

3) Apply Custom PCA

I. Standardize the dataset: $X_{std} = X - \mu/\sigma$, where X is the original data, μ is the mean and σ is the standard deviation statistical error that separates an observation from its predicted value.

II. Compute Covariance Matrix: $\Sigma = \frac{1}{n-1} X_{std}^T X_{std}$, where Σ is the covariance matrix, X_{std}^T is the transpose of the standardized data, and n is the number of observations.

III. Obtain Eigenvectors and Eigenvalues: $\Sigma v = \lambda v$, where v is the eigenvectors and λ is the eigenvalues.

IV. Select Top k Eigenvectors: Sort eigenvalues in descending order and select the top k eigenvectors corresponding to the largest eigenvalues. Construct the projection matrix W from these eigenvectors.

V. Transform Data: $Y = X_{std} W$, Where Y is the transformed data in the new k-dimensional space.

4) Perform Regression with PCA optimized data:

I. Split PCA-transformed data into training (80%) and testing (20%) sets.

II. Apply the following 7 models in the dataset Y with 2 component PCA, 3 component PCA and 5 component PCA: MLR, RR, LR, ELR, SVM, Vanilla ANN and Custom DNN

III. Evaluate performance for each configuration

5) Compare Results: Compare performance metrics from initial and PCA-transformed datasets to get the improved results.

3.7.1. Proposed Model

For this work, five regression model, and two neural network-based architecture models are introduced.

3.7.1.1. Multilinear Regression

Multilinear regression (MLR) is an analytical method of predicting the value of a specific metric by taking into account several independent variables. This algorithm is used to find the optimal straight line to predict the AQI value for various parameters. First, we define the independent features as x1, x2, ……., x10. Then, we have defined the dependent feature as y i.e., AQI. It can be mathematically represented via equation 1.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n + \epsilon \quad (1)$$

For nth observation, β_0 is a constant term also known as y-intercept and β_n is the slope coefficient for each explanatory variable. Parameter ϵ stands for the statistical error that separates an observation from its predicted value.

3.7.1.2. Ridge, Lasso and Elastic Net Regression

It is common for real-time unbalanced datasets to have a significantly larger number of input variables than observations. With many predictors, fitting the full model without penalization will result in large prediction intervals. For that reason, we have used three model tuning regression method i.e., Ridge Regression (RR), Lasso Regression (LR) and Elastic Net Regression (ELR) to analyze the problem better. RR constitutes a model refinement technique employed for the analysis of datasets afflicted by multicollinearity [15]. Employing L2 regularization, this method addresses situations where the occurrence of multicollinearity imparts bias to least-squares estimates and induces elevated variances. The mathematical representation of RR is in equation 2.

$$RRL_2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \quad (2)$$

Here, the L2 term is equivalent to the square of the coefficient magnitudes (βj) and the regularisation penalty is denoted by λ. As we increase the value of λ this constraint causes the value of the coefficient to tend towards zero [16]. By applying a penalty proportional to the absolute value of the magnitude of the coefficients, Lasso regression carries out L1 regularisation (equation 3). Certain coefficients may become zero and be removed from the model as a result of this kind of regularisation, producing sparse models with few coefficients.

$$LRL_1 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \quad (3)$$

To improve the model prediction rate, both regularization of lasso and ridge are combined to produce the elastic net model loss. The collinearity coefficient is difficult to eradicate, according to the fundamental rationale behind. In this study, the Grid Search approach (GS) for hyperparameter tweaking has been employed to enhance the model's performance. This research makes use of the Scikit-Learn class GridSearchCV. The GridSearchCV evaluates, all possible combinations of parameter values and finally, the best parameter combination is retained. After tuning the optimal λ value is 1, 0.0001, and 0.0001, for RR, LR, and ELR, respectively.

3.7.1.3. Support Vector Machine (SVM)

SVM divides data into multiple classes by locating a hyperplane in a high-dimensional space [17]. The primary objective is to optimize the margin, which is the space between the hyperplane and the closest support vectors, while simultaneously minimizing the error SVM uses a technique called the kernel trick to translate the feature space into a higher-dimensional space in order to improve data point separation when dealing with non-linearly separable data [18], [19]. The kernel functions used in the SVM model are linear and Radial Basis Function (RBF). The hyperparameter C, and epsilon, are taken as 5, and 1, respectively after tuning for the proposed model. C is the regularization parameter that controls the trade-off between achieving a low training and testing error while maximizing the margin. Epsilon is the width of the epsilon-insensitive tube in SVM Regression. It defines a margin of tolerance where no penalty is given to errors.

3.7.1.4. Artificial Neural Network (ANN)

The best model for automatically identifying and simulating intricate non-linear correlations between the "output" (i.e. AQI) and the "inputs" (i.e. NO, NO2, NH3 etc.) of the network is the artificial neural network (ANN), which can also take into account all potential interactions between the input variables [20]. As imbalanced dataset contains a large proportions of skewed classes so conventional regression method fails to recognise the non-linear relationship between inputs and outputs. The architectural depiction of the ANN model employed in this study is displayed in Figure. 2.

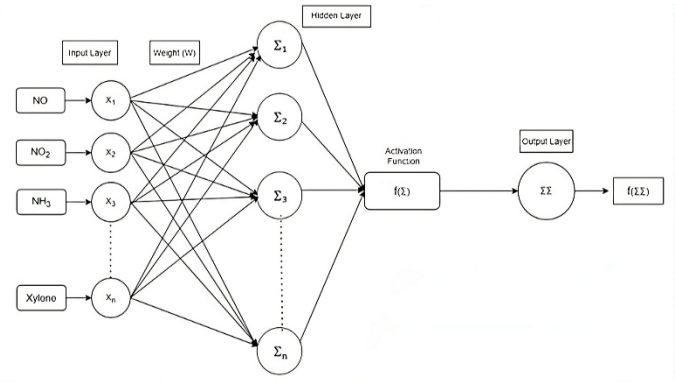


Figure 2: Visual Representation of ANN Architecture

The ANN consists of one input, one hidden and one output layer. The input features are represented as X1, X2, X3,…., Xn. The synaptic weights between the input and hidden layer are defined as W1, W2,…Wn. The biases are taken as b0 and b1 for hidden and output layer respectively. Σ denotes each node takes the weighted sum of its inputs, and passes it through a non-linear activation function. The mathematical equation of tanh is represented via equation 4.

$$f(\Sigma) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} = \frac{1-e^{-2x}}{1+e^{-2x}} \quad (4)$$

The AQI value is then predicted by feeding the hidden layer's output into the output layer. To reduce the error rate during backpropagation, the gradient descent is calculated and used to modify the weights and biases. The ANN model has been analysed combined with PCA for 2, 3 and 5 components respectively to increase the model interpretability and computational efficiency.

3.7.1.5. Deep Neural Network (DNN)

DNN refers to a multilayer perceptron model that has more than one hidden layer. When it comes to assessing the AQI value, DNN is a more potent and reliable neural network model than ANN. DNN exhibits of layer wise feature extraction methodology and combine low level feature to generate high level features [21], [22]. Figure. 3 shows the visual representation of DNN model. It is composed of one input layer with 10 neurons, two hidden layer (i.e., 256 and 512 neurons respectively) and one output layer. Weights and biases are initialised to the proper values at each layer of the network. Rectified Linear Unit (ReLu) activation function (represented by equation 5) and linear function are utilised at the hidden layer and output layer, respectively. To achieve state of the art performance, the model has then been evaluated using a combination of PCA and DNN.

$$\text{ReLu}\ (z) = \max\ (0, z) \quad (5)$$

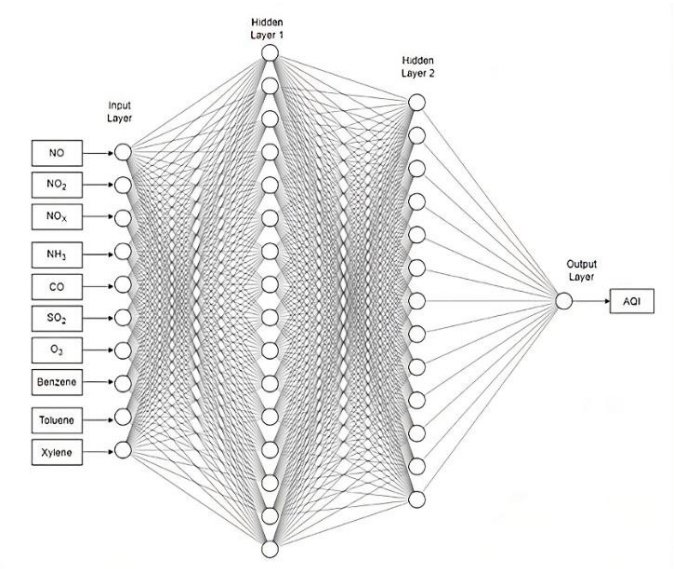


Figure. 3: Deep Neural Network Model Structure.

4. Results

The models for forecasting the AQI level based on different atmospheric pollution characteristics are

empirically evaluated in this section. Metrics such as mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R2) have been used to evaluate the model. Every model has been simultaneously trained, examined, and integrated using PCA and non-PCA analysis. For RR, LR, ELR, and SVM regression, the Grid Search method has been used.

Table 3: Comparison of Regression model results without PCA.

Model	Non PCA			
	Normal		GS	
	MSE	R2	MSE	R2
MLR	440.823	0.752		
RR	439.022	0.753	439.022	0.753
LR	446.132	0.749	439.294	0.753
ELR	510.646	0.712	439.148	0.753

The performance characteristics of a standard regression model with and without Principal Component Analysis (PCA) integration are shown in Tables 3 and 4, respectively. We can clearly see that RR has outperformed as compared to other model in terms of MSE and R2 score.

MLR and LR has slightly degraded performance in both normal and grid search method (GS) as compared to RR. After using PCA, it is shown that using five principal components produces outcomes that are almost similar to those that come from non-PCA analysis.

The association between the independent variables and AQI is explained in great detail by the trained MLR model. Equation 6 provides a statistical illustration of the concept represented in equation 1. Because LR regression employs the L1 norm, some of the coefficients are absolutely zero. This shows that the relevant elements have been effectively excluded from the model, rendering them useless in predicting the target variable. This increases model interpretability while also possibly improving prediction performance by decreasing overfitting.

$$MLR(Y_{AQI}) = [4.6,\ 27.41,\ -18.07,\ 10.84,\ 4.02,\ -0.21,\ 18.98,\ 4.85,\ 1.61,\ -0.27]_{1x10} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix}_{10x1} + 92.57 \quad (6)$$

The equation 6 represents the coefficient matrix (β_1= 4.6, $\beta_2 = 27.41$, ……., $\beta_{10} = -0.27$), independent features from NO to Toluene (X1, X2,...,X10)and intercept value($\beta_0 = 92.57$) for MLR. The values from β_0 to β_{10} are experimentally obtained corresponding to the MLR entry of table 3 without PCA. Similarly, equation 7, equation 8 and equation 9 defines the coefficient matrix, independent features and intercept values used for RR, LR and ELR model respectively. The statistical equations from 7 to 9 are generated with all the features without using PCA. The final combined PCA model equations along with coefficient matrix and intercept can also be generated in the similar way.

$$RR(Y_{AQI}) = [2.01,\ 21.07,\ -10.47,\ 10.87,\ 3.99,\ -0.19,\ 18.95,\ 4.81,\ 1.62,\ -0.27]_{1x10} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix}_{10x1} + 92.57 \quad (7)$$

$$LR(Y_{AQI}) = [-1.47e\text{-}02,\ 1.15e+01,\ 0,\ 1.11e+01,\ 3.28,\ 0,\ 1.8e+01,\ 4.46,\ 8.99e\text{-}01,\ 0]_{1x10} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix}_{10x1} + 92.57 \quad (8)$$

$$ELR(Y_{AQI}) = [-1.41,\ 5.07,\ 3.23,\ 9.27,\ 4.36,\ 2.07,\ 12.3,\ 3.81,\ 2,\ 0.97]_{1x10} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix}_{10x1} + 92.57 \quad (9)$$

As compared to traditional regression model SVM works better than in terms of R2 score. Table 5 describes that RBF with GS approach has the lowest RMSE and R2 score as compared to linear and normal RBF. With implementation of combined PCA and SVM, the model able to able to outperformed in each type of kernel by high margin. In regression, accuracy refers to the model's ability to forecast the percentage difference between the actual and estimated values. Table 5 compares all three types of SVM kernels to determine the overall best accuracy.

The derived equations for the linear and RBF kernels are provided in equations 10 and 11, respectively. The results of the AQI prediction are obtained using these equations. For linear SVM, positive coefficients indicate that a factor has a positive effect on predicted AQI, while negative coefficients indicate a negative effect. The intercept term represents the baseline AQI value. In contrast, with the RBF-kernelized SVM, the dual coefficients indicate the importance of the support vector in defining the decision region, with positive and negative coefficients indicating the direction of the effect on the predicted AQI.

Table 4: Comparison of Regression Model results with PCA

Model Name	PCA				Model				Type			
	2 PCA Component				3 PCA Component				5 PCA Component			
	Normal		GS		Normal		GS		Normal		GS	
	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2
MLR	546.891	0.692			461.973	0.740			450.89	0.746		
RR	546.896	0.692	546.997	0.692	461.979	0.740	462.039	0.74	450.90	0.746	450.97	0.746
LR	549.518	0.691	546.909	0.692	467.196	0.737	462.006	0.74	457.81	0.742	450.89	0.746
ELR	596.489	0.664	547.008	0.692	527.029	0.703	461.988	0.74	520.04	0.707	451.07	0.746

Table 5: SVM Regression Results with PCA and NON PCA.

Kernel Type	NON PCA			PCA			Model			Type		
				2 PCA Component			3 PCA Component			5 PCA Component		
	RMSE	R2	Accuracy	RMSE	R2	Accuracy	RMSE	R2	Accuracy	RMSE	R2	Accuracy
Linear	21.476	0.725		23.619	0.686		21.667	0.736		21.383	0.742	
RBF	20.867	0.741	80.86	20.540	0.762	80.16	20.351	0.767	80.523	20.217	0.770	81.00
RBF_GS	20.045	0.761		20.186	0.770		19.541	0.785		18.984	0.797	

Table 6: Comparison of ANN Model Results.

Metrics	NON PCA	PCA Model Type		
		2 PCA Component	3 PCA Component	5 PCA Component
MAE	14.616	15.025	14.979	15.235
MSE	390.485	408.578	398.932	427.794
RMSE	19.760	20.213	19.973	20.683
MAPE	0.198	0.206	0.205	0.207
R2	0.757	0.746	0.752	0.734
Accuracy	80.152	79.341	79.442	79.206

After training using linear SVM equation 10 shows the final weight matrix (W1= -0.3, W2= 0.52,……, W10= 0.09), input feature matrix (X1, X2,...,X10) and bias as 9.58 for computing the AQI output. The RBF kernel based SVM represented via equation 11 analyze for the weights consisting between -1 and 1, bias term (113.14), input feature vectors from X1 to X10 and each support vectors $(K(X,X_1),K(X,X_2),\ldots\ldots\ldots\ldots,K(X,X_{10}))$ passing through X1 to X10.

F(AQI) = [-0.30, 0.52, 0.03, 1.99, 4.41, 0.21, 0.59, 1.28, 0.18,

$$0.09]_{1x10}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{10} \end{bmatrix}_{10x1} + 9.58 \qquad (10)$$

Table 6 shows the result after training the imbalanced dataset using ANN architecture. ANN outperformed all of the earlier models in terms of strength and performance. It achieved lowest RMSE and MSE as compared to SVM, RR, LR, ELR and MLR. ANN Model

Table 7: Comparison of DNN Network Results

Metrics	NON PCA	PCA Model Type		
		2 PCA Component	3 PCA Component	5 PCA Component
MAE	13.08907	15.67781	17.27715	15.70377
MSE	334.054	438.4483	583.0541	469.349
RMSE	18.27715	20.93916	24.14651	21.66446
MAPE	0.172188	0.213611	0.229836	0.208632
R2	0.792717	0.72794	0.638211	0.708766
Accuracy	82.78118	78.63886	77.01642	79.13676

obtained with ten inputs, one hidden layer with one twenty-eight neurons and final output layer with predicted AQI value.

$$f(AQI) = [-1\ -1\ -1\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1\ -1\ 1]_{1x10}\ [\ K(X,X_1)\ K(X,X_2)\ \vdots\ K(X,X_{10})\ ]_{10x1} + 113.14 \quad (11)$$

The DNN Model is a very sophisticated and potent neural network designed to manage unbalanced datasets. Table 7 shows that the DNN model works better than the ANN by significantly increasing the R2 score and simultaneously decreasing metrics like MSE, MAE, and RMSE. Notably, the DNN model gets an accuracy rate of 82.78%, outperforming all other models in the comparison

A lower MSE and a higher R^2 score indicate that the DNN model also performs even better than conventional regression models. The DNN model is more successful without requiring feature extraction, as seen by its lower RMSE and greater accuracy when compared to the SVM model using five component PCA. During training, ANN and DNN possess the ability to autonomously extract meaningful features from raw data. They can adapt to high-dimensional input spaces without explicit feature extraction. ANN and DNN model with combined PCA result has been shown in table 6 and table 7 respectively. However, complicated connections among features can be handled by the proposed DNN model without the need for PCA. This not only increases efficiency and interpretability, but also reduces training time and computational cost, making the DNN model a state-of-the-art approach for various performance metrics

DNN model takes ten neuronal inputs, and two hidden layers consisting of two fifty-six and five hundred twelve neurons respectively. The output layer involves a 512×1- dimensional weight matrix along with a linear function to derive the AQI value.

Figure 4 and 5 illustrate the rate of MSE change for the ANN model employing 10 input features and utilizing dimensionality reduction with 2 features, respectively. It is clear that the ANN model with 10 input features converges much more quickly than the one using 2 features plus PCA because of inadequate feature representation. Similarly, Figure 6 and 7 depict the MSE convergence dynamics during the training of neural network architectures for final output layer prediction. The DNN model exhibits standard MSE performance in both training and testing datasets, while the DNN utilizing 2 principal components tends to converge prematurely, resulting in less accurate predictions for the final output layer.

Consequently, our original DNN synaptic model architecture demonstrates superior performance with robust convergence compared to alternative networks.

DNN models can converge faster and produce more accurate results without the usage of PCA. The suggested DNN model is advantageous for achieving optimal outcomes, especially in scenarios involving imbalanced datasets.

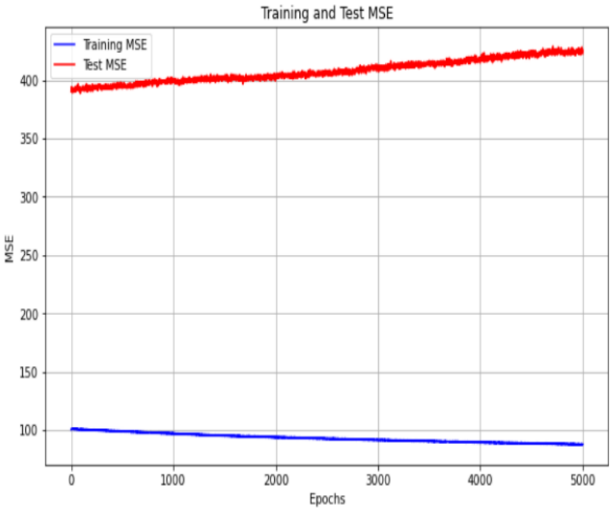


Figure 4: ANN MSE Convergence.

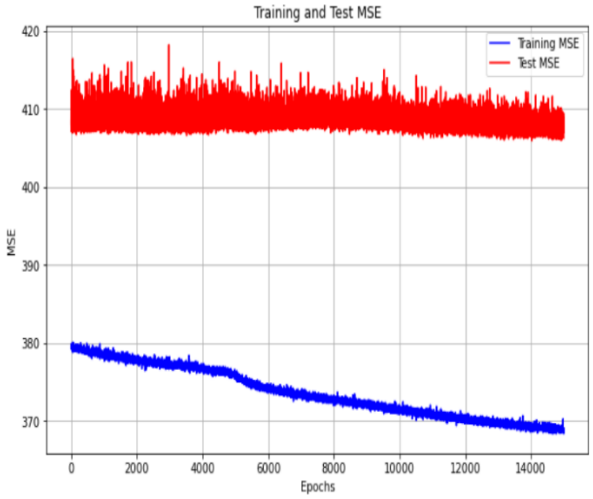


Figure 5: ANN 2PCA MSE Curve

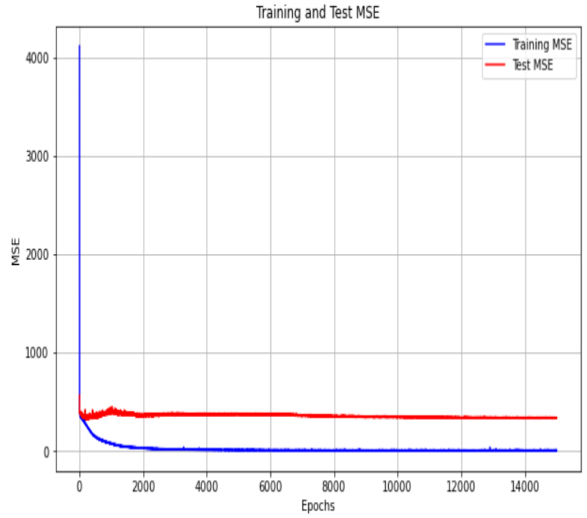


Figure 6: DNN MSE Curve

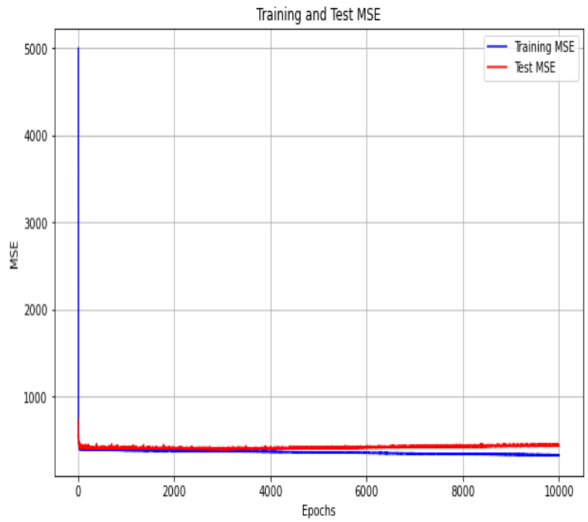


Figure 7: DNN 2PCA MSE Plot

PCA effectively reduces the dimensionality of the dataset into 2, 3, and 5 components, respectively. It identifies the most significant features by projecting the data into k-dimensional spaces that capture the most critical variance. This process streamlines the data, minimizing noise and redundancy, which makes the data more regular and structured. By introducing the most relevant components, PCA allows the DNN to operate more efficiently, avoiding overfitting and improving its ability to learn complex patterns in the data. This enhancement in data regularity enables the DNN to outperform other regression models in terms of performance and interpretability.

5. Conclusion

Though "Accuracy" is a prime indicator of model performance however its reliability depends upon the dataset. In an out of balanced dataset, the majority class makes up a large portion of the training dataset, while the minority class is underrepresented in the dataset. The problem with a model trained on this out of balanced data set is that the model learns that it can achieve high accuracy by consistently predicting the majority class, even if recognizing the minority class is equal or more important when applying the model to a real-world scenario. Dealing with real-time out of balance datasets presents substantial difficulties, especially in jobs such as anticipating the air top quality index as well as creating appropriate regression network designs. The complexity stems from the dynamic character of the environment, the variability of pollutant levels, and their geographical and temporal irregularity. In the present work, we utilize an air pollution dataset comprising various pollutants to address this challenging problem. The dataset is extensively looked into making use of exploratory information evaluation strategies to efficiently pre-process it for research study objectives. As a result of large number of parameters in the dataset, dimensionality reduction methods such as PCA are utilized. MLR, RR, LR, ELR, and SVM are the machine learning algorithms used for training and testing with the goal of predicting the air quality index (AQI). Additionally, neural network designs particularly ANN along with DNN are checked out for more evaluation.

Each model is trained with and without PCA to compare error rate and overall performance. ANN design shows durable convergence and also exceptional efficiency as contrasted to different ML design. By harnessing the power of high-level feature extraction, the DNN model is able to surpasses traditional regression techniques and even outperforms other neural network designs like ANN. This study showcases the efficacy of neural networks, particularly the DNN model that exhibits superior accuracy, lower MSE, and higher R2 scores, indicating its ability to automatically identify relevant characteristics from unprocessed data, eliminating the requirement for explicit feature extraction methods.

6. Future Work

Future research may enhance the management of imbalanced datasets by utilizing sophisticated neural architectures such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Additionally, real time imbalance may also be improved by using ensemble methods and transfer learning. Furthermore, statistical time series models like Seasonal Autoregressive Integrated Moving Average (SARIMA), Autoregressive Integrated Moving Average (ARIMA), and Long Short-Term Memory (LSTM) can forecast trends for dangerous pollutants including CO, NO, and NO2 etc.

Acknowledgement

We express sincere gratitude towards the Department of Electronics and Communication Engineering, Heritage Institute of Technology, Kolkata and our parents for providing the research environment and support that enabled us to undertake this research work.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. D. Pozzolo, O. Caelen, Y.A.L Borgne, S. Waterschoot, G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert Syst. Appl.* 2014, 41, 4915–4928, doi: https://doi.org/10.1016/j.eswa.2014.02.026.

[2] B. Anuradha, V. C. Veera Reddy, "ANN for classification of cardiac arrhythmias," *Asian Research Publishing Network Journal of Engineering and Applied Sciences*, vol.3, no.3, 1-6, 2008.

[3] L. Bruzzone, S. B. Serpico, "A classification of imbalanced remote-sensing data by neural networks," *Pattern Recognition Letters*, vol.18, pp.1323-1328, 1997, doi: https://doi.org/10.1016/S0167-8655(97)00109-8.

[4] G. H. Nguyen, A. Bouzerdou, S. L. Phung, "A supervised learning approach for imbalanced data sets," *Proc. of the 19th International Conference on Pattern Recognition*, 1-4, 2008, doi: 10.1109/ICPR.2008.4761278.

[5] G. Pang, C. Shen, L. Cao, A. Van Den Hengel, , "Deep learning for anomaly detection: A review," *ACM Comput. Surv. (CSUR)*, 54, 38, 2021, doi: https://doi.org/10.1145/3439950.

[6] A. Adam, M. Shapiai, Z. Ibrahim, M. Khalid, "A Modified Artificial Neural Network Learning Algorithm for Imbalanced Data Set Problem," *International Conference on Computational Intelligence, Communication Systems and Networks*, CICSyN 2010, doi: 10.1109/CICSyN.2010.9.

[7] S. Wang, W. Liu, J. Wu, "Training Deep Neural Networks on Imbalanced Data Sets," *International Joint Conference on Neural Networks (IJCNN)*, 4368-4374, 2016, doi: 10.1109/IJCNN.2016.7727770.

[8] A. Li, S. Liang, A. Wang, J. Qin, "Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques," *American Society for Photogrammetry and Remote Sensing*, Vol. 73, No. 10, 1149–1157, 2007, doi: 10.14358/PERS.73.10.1149.

[9] Y. Tang, V. N. Chawla, "SVMs Modeling for Highly Imbalanced Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 39, 281 – 288, 2008, doi: 10.1109/TSMCB.2008.2002909.

[10] Y. Yang, K. Zha, "Delving into Deep Imbalanced Regression," *ICML 2021*, https://arxiv.org/abs/2102.09554.

[11] C. Huang, Y. Li, C. L. Change, X. Tang, "Learning deep representation for imbalanced classification," *IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016, doi: 10.1109/CVPR.2016.580.

[12] M. Steininger, K. Kobs, P. Davidson, "Density-based weighting for imbalanced regression," *Mach Learn*, 110, 2187–2211, 2021, doi: https://doi.org/10.1007/s10994-021-06023-5.

[13] A. Rahim, N.A. Rashid, A. Nayan, A. Ahmad, "SMOTE Approach to Imbalanced Dataset in Logistic Regression Analysis," *ICMS 2017*, 429-433, 2019, doi: https://doi.org/10.1007/978-981-13-7279-7_53.

[14] P. Branco, L. Torgo, P. R. Ribeiro, "SMOGN: a Pre-processing Approach for Imbalanced Regression," *In First international workshop on learning with imbalanced domains: Theory and applications*, pages 36–50. PMLR, 2017.

[15] C. Peng, Q. Cheng, "Discriminative Ridge Machine: A Classifier for High-Dimensional Data or Imbalanced Data," *IEEE Trans. on Neural Networks and Learning Systems*, 2595 – 2609, 2020, doi: 10.1109/TNNLS.2020.3006877.

[16] A. SzeTo, K. C. Wong, "A Weight-Selection Strategy on Training Deep Neural Networks for Imbalanced Classification," *International Conference Image Analysis and Recognition*, 3-10, 2017, doi: https://doi.org/10.1007/978-3-319-59876-5_1.

[17] R. Akbani, S. Kwek, N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," *European Conference on Machine Learning (ECML)*, 39–50, 2004, doi: https://doi.org/10.1007/978-3-540-30115-8_7

[18] Y. H. Liu, Y. T. Chen, S. S. Lu, "Face Detection Using Kernel PCA and Imbalanced SVM," *International Conference on Natural Computation*, 351–360, 2006, doi: https://doi.org/10.1007/11881070_50.

[19] J. Mathew, M. Luo, C. K. Pang, H. L. Chan, "Kernel-based smote for SVM classification of imbalanced datasets," *IECON*,.1127-1132, 2015, doi: 10.1109/IECON.2015.7392251.

[20] R. Anand, K. G. Mehrotra, C.K. Mohan, S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets;" *IEEE Trans. Neural Networks* 4, 962–969, 1993, doi: 10.1109/72.286891.

[21] H. Larochelle, Y. Bengio, J. Louradour, J. Lamblin, "Exploring strategies for training deep neural networks," *Journal of machine learning research*, vol 10, 1-40, 2009, doi: 10.1145/1577069.1577070.

[22] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, R. Togneri, "Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data," *IEEE Trans. Neural Network Learn System*, pp 3573 - 3587, 2017, doi: 10.1109/TNNLS.2017.2732482.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/)



SWARUP KUMAR MONDAL is a Software Engineer at CoreLogic India. He is an Electronics and Communication Engineering Undergrad from Heritage Institute of Technology, Kolkata, India. He has one research paper publication.

His research areas are Machine Learning, Deep Learning, Optimization and Medical Image Processing.



ANINDYA SEN is a Professor at the department of Electronics and Communication Engineering, Heritage Institute of Technology, a private autonomous engineering college in Anandapur, Kolkata, India. He received his B.E. from Jadavpur University, India in the year 1980, PhD from University of Minnesota, Twin Cities in 1996, and got his Post-Doctoral training from University of Chicago from 1996 to 2000. He currently holds one US patent and 65, research paper publications. His research interests include, Medical Image processing, Internet of things, Artificial Intelligence and VLSI design.

Received: 21 August, 2024, Revised: 28 September, 2024, Accepted: 29 September, 2024, Online: 11 October, 2024

DOI: https://doi.org/10.55708/js0310002

Conceptual Business Model Framework for AI-based Private 5G-IoT Networks

Laurence Banda*

Wits Business School (WBS), University of the Witwatersrand, Johannesburg, 2050, South Africa

*Corresponding author: Laurence Banda, +27 834282216, laurencebandad@gmail.com

ABSTRACT: The fusion of fifth generation (5G) networks, Internet of Things (IoT) and artificial intelligence (AI), referred to as intelligent connectivity by most industry experts, can be seen as a crucial success factor for sustainable digitalization. Until recently, research into these key triad technologies has been conducted in isolation. One of the promising applications of intelligent connectivity is wireless private networks. This article presents a conceptual business model framework that can be adopted by AI-driven private 5G-IoT networks. These emerging and disruptive private networks will certainly change the mobile business landscape. Private network operators should therefore rethink and adapt their business models in order to remain economically competitive, create innovative mobile services and support sustainable practices.

KEYWORDS: 5G, Artificial Intelligence (AI), Business Model, Conceptual Framework, Internet of Things (IoT), Private Networks

1. Introduction

Fifth generation (5G) networks are being widely deployed around the world with the hope that these cutting-edge mobile systems will contribute to the sustainable digital transformation and economic prosperity of society [1]. With the introduction of more affordable devices and the expansion of commercial 5G mobile networks, customer and business interests in the market are increasing rapidly. Furthermore, the Internet of Things (IoT) and artificial intelligence (AI), which are revolutionizing our modern world through digitalization, have been enabled in large part by 5G networks [2].

IoT technologies pervasively connect natural and artificial objects and people across all network infrastructures [3]. Furthermore, IoT is considered a multidisciplinary field that encompasses numerous areas, including natural ecosystems, infrastructure, public services, social activities, technology, and the business and economic sectors [4]. The aim is to create innovative network architectures, applications and services based on smart IoT networks and smart sensors based on the 5G-IoT technology standard. 5G-IoT is therefore expected to generate massive amounts of sensor data, which requires machine learning (ML) to analyze and make intelligent decisions.

AI is a relatively new technical science that studies theories, techniques, tools and application systems that mimic and enhance human intelligence [5]. In particular, through massive data learning, machines mimic human thinking and behavior. AI is a remarkable technology with applications that cannot be fully covered by traditional methods as current solutions require advanced automation and optimization. Mobile networks in particular generate large amounts of data and make them amenable to machine learning (ML) optimization techniques. Therefore, the application of AI to networks actually refers to the use of AI/ML methods to develop supporting functions for network optimization, configuration and management operations.

The foundation of "intelligent connectivity," as some have termed it, is the combination of 5G, IoT and AI [2,6]. This is the beginning of an era characterized by ubiquitous hyper-connectivity and highly contextualized and personalized experiences. Almost every aspect of our everyday lives could be affected, including the way we work and interact with colleagues, as well as the way we

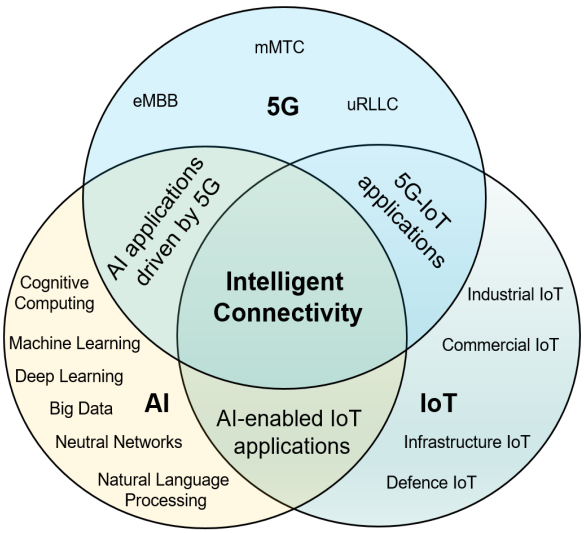


Figure 1: Fusion of 5G, IoT and AI for intelligent connectivity

consume digital entertainment [7]. Consequently, the convergence of 5G, IoT and AI will improve the quality of goods and services, lead to significant productivity gains and enable more effective use of global resources by both businesses and consumers. Key applications of intelligent connectivity include industrial IoT (iIoT), sustainable smart cities, Industry 4.0, and network security [2,6,7]. Figure 1 shows the fusion of 5G, IoT and AI for intelligent connectivity.

The 5th Generation Partnership Project (5GPP) has standardized private networks to provide specialized wireless connectivity with improved communication qualities, optimized services and tailored security features in specific locations [8]. These wireless local area networks are deployed in isolated and remote areas that are not adequately covered by traditional public cellular networks. 5G, IoT and AI are arguably the three most revolutionary technologies of the current decade. The three technologies were used independently until recently and each has exceeded our expectations. It should come as no surprise that adoption or integration of 5G, IoT, and AI is expected to increase as more businesses embrace the data trend. In addition, the combination of 5G, IoT and AI technologies has increased the number of network partners. To remain sustainable and profitable, AI-based private 5G-IoT network operators must therefore realign their business models.

This study contributes to the development of innovative business model solutions for novel triad technologies (5G, IoT and AI). In particular, the paper aims to contribute to research, development and innovation in AI-based private 5G-IoT networks by proposing a conceptual business model framework that can be adopted by operators of such networks. The article also presents some considerations for the cross-sector implementation of business models. Furthermore, the article is a continuation of our research publications on business models for 5G mobile network operators (MNOs) and private 5G-IoT networks [9,10]. The proposed business models are viewed from both a strategic and technology theoretical perspective. The rest of the article is structured as follows. Section II highlights related works from the literature. Section III discusses private 5G-IoT networks with AI capabilities. Section IV presents the proposed conceptual business model framework, while Section V discusses industrial applications of the proposed conceptual framework. Section VI concludes the article.

2. Related Work

Recent developments in 5G wireless networks, IoT and AI have led to attempts to review relevant literature. Specifically, [11] essentially conducted a techno-economic analysis of private 5G network architectures. To accelerate the achievement of return on investment (RoI) by private network operators and other third parties, the paper focuses primarily on cost elements associated with the network deployment of private 5G networks. However, the paper ignores other crucial elements of the AI-based private 5G IoT network value chain, such as the value proposition, customer interface and infrastructure management. In [10] new business models for private 5G-IoT networks were presented. The study first develops a conceptual framework for the private 5G-IoT network value chain and then describes several business models relevant to different use cases and vertical markets. Nevertheless, the authors' study did not address the AI component.

A theoretical framework for Internet of Everything (IoE) applications for creativity in the context of 5G and 6G ecosystems was introduced in [12]. The study discusses issues with the IoE-based business platform. This is particularly about creative, cross-domain business solutions that cover different application paradigms and business platforms. However, the authors of [12] did not identify the industry's vertical target customers. Business model options for 5G and future mobile network operators were discussed in [10,13]. The paper first outlines the limitations of current business models for mobile network operators, before presenting new business models for 5G and future networks that mobile network operators should consider when introducing such networks. There was more focus on traditional or public 5G networks, while less attention was paid to private 5G networks. A business model agenda for 5G and 6G players is discussed in [14]. The authors' work focuses on the telecommunications ecosystem and emphasizes the significant changes in business models brought about by 5G. In addition, potential future business models beyond 5G and 6G are identified and attention is drawn to the respective business opportunities of the players. However, important new technologies such as IoT and AI enabled by 5G and 6G networks were not taken into account.

This article differs from previous related works in that it exclusively addresses business models for AI-powered private 5G- IoT networks using a conceptual framework; it highlights some of the key concerns of private network operators related to certain AI-driven 5G-IoT business models; and it highlights some of the economic benefits and concerns associated with integrating 5G, IoT and AI technologies.

3. AI-based Private 5G-IoT Networks

The introduction of 5G operational networks, which are almost mature, is followed by the concept and architectural design of AI-powered private 5G-IoT networks. The GSMA estimates that around 65% of the world's population will have access to 5G and high-speed mobile internet connectivity by the end of 2025 [9]. Proper deployment of new technologies such as sensor-based and AI-powered IoT, integrated with the 5G Machine Type Communication (MTC) ecosystem and associated smart industries, will enable the growth of new smart industries and drive digitalization.

3.1. 5G-IoT-AI Integration

By integrating AI into 5G IoT networks, millions of simultaneous intelligent connections are expected across a range of smart devices, automated machines, connected homes, smart grids, and smart transportation systems [9]. The idea behind intelligent connectivity is to use 5G, IoT and AI together to accelerate technological advancement and open up new, disruptive digital services [7]. In the context of intelligent connectivity, artificial intelligence (AI) technologies will analyze, contextualize and present the digital data collected by the machines, devices and sensors of the 5G-IoT networks to users in a more meaningful and useful way [6]. This would facilitate the delivery of personalized experiences to users and improve decision-making, leading to richer and more satisfying interactions between individuals and their environment.

Intelligent connectivity aims to achieve several key goals, including enhanced communication reliability, higher data rates, multiple sensor connectivity, and incredibly low power consumption [15]. However, the integration of 5G, IoT and AI faces implementation challenges due to different platforms, including: device platform (e.g. radio frequency security and device battery life), application platform (e.g. centralization constraints), and the central network platform (e.g. enabling always-online services) [15]. By dividing the physical network infrastructure into multiple logical networks (slices), each with unique service requirements and performance standards, network slicing technologies can help resolve these issues [9,16].

3.2. The Concept of AI-driven 5G-IoT Networks

A private 5G network, also known as a non-public 5G network, is a local area network that provides dedicated wireless connectivity within a specific area. Its owner can manage the network independently and has complete control over all aspects of the network, including resource scheduling, security, and priority access [17]. Therefore, an AI-based private 5G-IoT network can be described as a non-public 5G network enabled by AI technologies and tailored to IoT-based vertical markets.

Private 5G networks can be deployed in either standalone (independent) or public network integrated (dependent) mode, depending on the availability of spectrum resources and infrastructure as well as the level of network management and access control. A standalone private 5G network is a network that is deployed as an isolated and independent system without relying on a public 5G network. Conversely, a private 5G network integrated into a public network is anchored in the public network and therefore has a lower level of customization, self-control and security compared to standalone mode [9,17]. Figure 2 shows typical private 5G networks supporting various vertical markets as illustrated in [18].

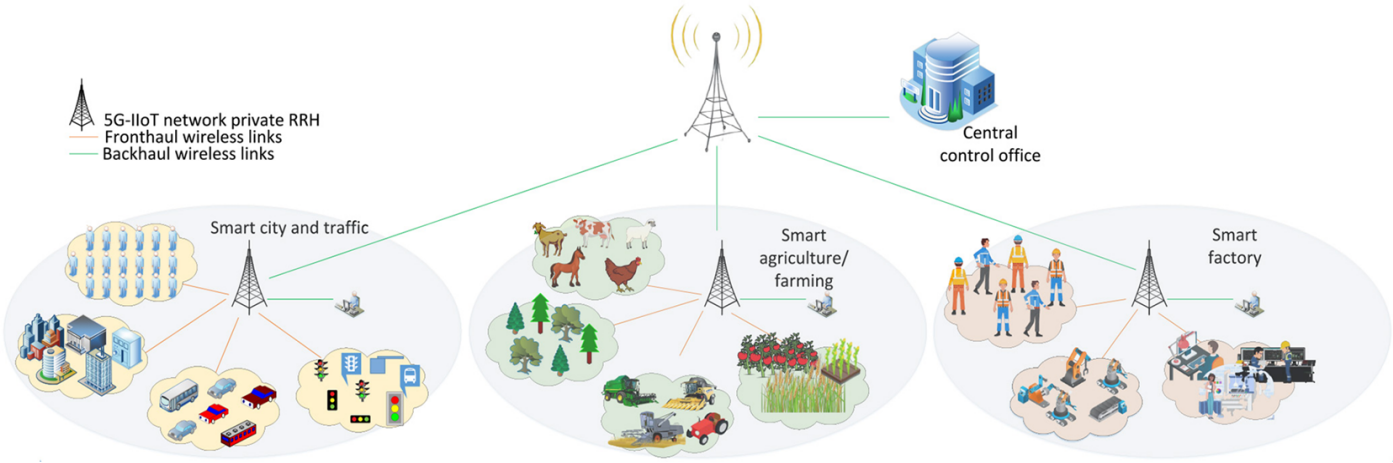


Figure 2: Private 5G networks [18]

4. Conceptual Business Model Framework

4.1. The Business Model Theoretical Concept

Since the mid-1990s, when the Internet first became widespread, the concept of business model (BM) has become increasingly popular and is now a key area of interest for researchers [19]. The term "business model" is widely used in academic literature and management practice, but there is no formal, accepted definition [19]. This is primarily because business models can be used in different research and practical fields [9]. A business model in the context of this study can be defined as follows:

The rationale for implementing the business plan of an AI-based private 5G-IoT network by leveraging interconnected elements such as customer relationships, value proposition, financial aspects and infrastructure management to create, deliver and capture value within the network ecosystem.

The present study adopts two distinct research perspectives to examine the business model concept. The first is a strategic perspective that focuses on the value proposition, creation, delivery, and capture activities within the AI-driven 5G-IoT value chain. The second perspective is technological in nature, wherein AI-driven 5G-IoT networks are considered as technological enablers for innovative business models aimed at private network operators [9]. Figure 3 illustrates the business model for AI-driven private 5G-IoT networks from both a strategic and technological perspective.

4.2. Proposed AI-based Private 5G-IoT Value Chain

Based on the business model ontology presented in [20], the relationships between the value chain components within the global market should be precisely defined in order to develop viable business models for AI-based private 5G-IoT networks. The following value chain components are proposed and shown in Figure 4.

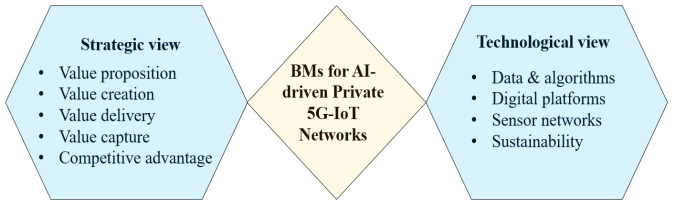


Figure 3: Strategic and technological views of AI-driven private 5G-IoT networks' business models

Who? Which target customer segments will benefit from the proposed value offered by the private network operator? This part of the value chain is about identifying the specific needs of these customers and offering tailored services to ensure their satisfaction. Customer segments include government and public institutions (such as public safety and smart cities), private companies (such as media and entertainment), industries (such as energy and healthcare), and individual consumers.

What? This value chain component explains the private network operator's value proposition for a specific customer group with the aim of generating added value for both the provider and the customer. These services include: enhanced mobile broadband (eMBB) applications such as ultra-high definition video; massive machine type communication (mMTC) applications such as smart factories; and ultra-reliable low latency communication (uRLLC) applications such as remote surgery.

How? This part of the value chain describes the technology design, network infrastructure and network resources available to the private network operator to provide the relevant services to intended customers. The following could be included in the "How?" component: AI-based key technologies for private 5G-IoT networks; sharing site infrastructure such as antennas, baseband resources, radio resources and tower space; spectrum sensing and acquisition such as mmWave band, sub-3GHz band and C-band; and collaboration with partner networks such as platform suppliers, device manufacturers and device providers.

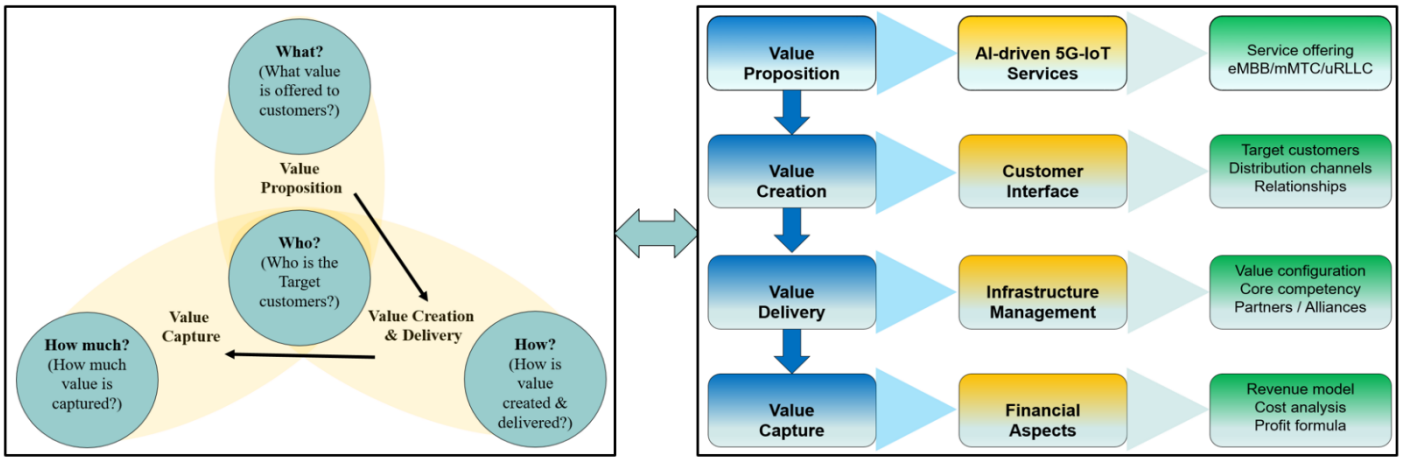


Figure 4: Relationships between value chain components of AI-based private 5G-IoT network

How much? This part of the value chain provides a blueprint for how an operator wants to maximize profits while minimizing costs within the AI-based 5G private IoT network ecosystem: How much value will the private network operator derive from the target customer segment? The "How much?" components include: revenue structures such as revenue streams, profit margins and return on investment (RoI); pricing strategies such as unit cost pricing and on-demand pricing; and cost structures such as deployment costs, capital investments and operational costs.

4.3. Proposed Conceptual Business Model Framework

A conceptual framework serves as an analytical tool that provides a preliminary theory of a phenomenon under study [21]. It establishes the connection between concepts by illustrating the theories that impact the research investigation, thereby providing a theoretical basis for designing and understanding the results [21]. Table 1 shows the ten components of the proposed conceptual business model framework for AI-based private 5G-IoT networks based on the business model canvas developed by [20].

Key partners: demonstrates the need for collaborative arrangements with external organizations to effectively deliver and monetize the proposed value.

Key activities: describes the activities and competencies required to implement the private network operator's business model.

Value proposition: provides a comprehensive overview of a range of products and services offered by a private network operator.

Customer segment: describes the customer segments to which a private network operator wants to offer value. These are the target groups of the value proposition.

Customer relationships: explains the type of connections that a private network operator establishes between itself and its various customer segments.

Key resources: these are the resources available to a private network operator to ensure that value is created and delivered in line with the operator's business objectives.

Distribution channels: describes various means that a private network operator can use to engage with its customers and offer them the value proposition.

Infrastructure management: describes the technology design, network infrastructure, and network resources available to the private network operator to deliver the relevant proposed value to intended customers.

Cost structure: sums up the monetary consequences of the resources used in the business model of a private network operator.

Revenue streams: describes the way a private network operator makes money through a variety of revenue flows.

5. Discussion of Industrial Applications of the Proposed Conceptual Framework

5.1. Sustainable Industry using AI-based Private 5G-IoT Networks, Business Models and Use Cases

Different business models can be used when implementing 5G cellular technologies for AI-driven private 5G-IoT network applications. The choice of business model is influenced by a number of factors, including the industrial sector, a specific use case, capital and operating costs, and the scalability of the network [9]. In addition to providing reliable services to key stakeholders, private 5G networks are fast, easy to manage, secure and can operate in remote or isolated locations. Table 2 summarizes typical business models, vertical industry target markets, key drivers and use cases for AI-based private 5G-IoT networks.

Furthermore, one of the key killer use cases identified in the recent COP28 climate agenda is the use of emerging digital technologies such as private 5G-IoT and AI systems to monitor natural ecosystems [9,10].

5.2. Business Model Implementation Considerations

(i) Adaptive service-based approach

Some of the more exciting 5G technologies such as network slicing [16] and digital platforms [12] require a variety of key players and partners to deliver a range of intelligent services to customers. Therefore, a shift from a product-based to a flexible service-based business model is required. This is crucial because customer needs can change over time. Adaptive, service-based business models have the potential to improve user satisfaction and increase revenue for partners and operators alike.

(ii) Security-focused business model approach

Research on 5G business models has focused more on the economic value aspects of key network performance metrics such as data throughput, latency, reliability and device connectivity density, with relatively little attention paid to security objectives such as confidentiality and privacy data integrity. However, the involvement of many role actors and the complexity of the 5G ecosystem, each with different security levels, increases the vulnerability to cyberattacks [4]. Consequently, new business models should be security-focused by integrating AI and IoT-based forensics as well as various other means of combating cyber threats to ensure confidentiality, privacy and data integrity. AI and IoT forensics are also suitable tools to avert impending quantum security risks on 5G and future networks [22].

Table 1: Proposed Conceptual Business Model Framework for AI-based Private 5G-IoT Networks

Key Partners	**Key Activities**	**Value Propositions**	**Customer Segment**	**Customer Relationships**
- AI algorithm developers - IoT sensor manufacturers - Public MNOs - Platform owners & designers - Regulations & policy makers	- Managing partnerships - Network & platform maintenance - Mobile service delivery - Customer technical support - Research & development (R&D)	- Immersive mobile services - Automated manufacturing - High-speed broadband access - Smart city services - Automated factory production lines	- Mass market - Niche market - Multi-sided platforms - Individual consumers	- Personal assistance - Self-service - Automated service - Co-creation partner services
Key Resources	**Distribution Channels**	**Infrastructure Management**	**Cost Structure**	**Revenue Streams**
- Physical - Human - Intellectual - Financial	- Web-based sales - Operator's own store - Partner's store - Wholesale	- Spectrum acquisition - Backhaul infrastructure - Radio equipment - Tower space management	- Fixed costs - Variable costs - Capital expenditure - Operational expenditure	- Subscription fees - Usage fees - Leasing/renting assets - Brokerage fees - Advertising

Table 2: AI-based Private 5G-IoT Network Business Models, Vertical Industries, Key Drivers, and Use Cases

AI- based Private 5G-IoT Network Business Model	**Vertical Industry**	**Key Drivers**	**Use Cases**
Smart Factory	• Manufacturing	• Proliferation of mobile robotics in process automation.	• Connected workers • Production line flexibility • End-to-end logistics
Smart Mining	• Mining	• Demand for high safety standards in the mining industry. • Rapid transition from human-driven machine control to automated remote controlled applications	• Early warning disaster signaling • Safety signal dissemination • Automated drilling machine communication
e-Health (Healthcare 5.0)	• Health	• Advancement in medical research • Low density of health practitioners • Upsurge in remote surgery practice	• Wearable health monitoring devices. • Smart implants for medical diagnostics • Connected contact lenses
Smart Sea Port	• Maritime	• Surge in sea port activities mainly driven by cargo transportation which cannot be handled by the aviation industry.	• Remote-controlled & automated cranes • Automated guided vehicles (AGVs) • Unmanned aerial vehicles (UAVs) • Sea condition monitoring
Smart Airport (Airport 4.0)	• Aviation	• Rise in air freight activities. • Upsurge of passenger air traffic, post COVID-19 era.	• Automated AI-based boarding procedures for passengers. • Constant security monitoring • Automated fever detection • Optimized luggage handing
Smart Railway	• Transportation	• Increase in use of railway systems as a means of inland cargo transportation. • Increased number of train commuters due to road congestion.	• Smooth scheduling & operations of trains. • Secure and low latency communication between drivers and signaling controllers. • Automated AI-based boarding procedures for passengers.
Smart City	• Public	• Rapid urban/city population explosion • Increase in CO_2 emission levels in urban areas	• Remote monitoring of roads and city infrastructure. • Automated water sanitation and waste management systems • Air pollution level detection
Smart Grid	• Energy	• Multiple wireless sensors with uninterrupted 5G connectivity deployed in specific substations.	• Effective monitoring and forecast of energy demand and consumption • Automated adjustment of electric profile via load control.
Smart Farming	• Agriculture	• Multiple wireless sensors with uninterrupted 5G connectivity deployed in specific crop fields.	• Optimized soil moisture content • Efficient use of water and fertilizers via customized applications
e-Learning	• Education	• Immersive real-time interaction via virtual presence applications with minimal visual or audio delays	• Immersive real-time virtual interaction between learners and educators. • Tele-education use cases for distance-learning courses.

(iii) Support for green and sustainable energy-efficient communication

Supporting green and energy-efficient communication technologies through business models for AI-based private 5G-IoT networks is critical to the environmental sustainability of these networks. To reduce the reliance on conventional fossil fuels to power network devices, green energy sources must be utilized. One of the goals of 5G-based network design is to minimize carbon emissions and reduce energy consumption of network devices without compromising network quality. In [23], research was conducted on energy harvesting techniques that enable communication devices to harvest energy from a variety of renewable sources. However, so far neither industry nor academia have shown much interest in exploring business models aimed at enabling environmental friendliness in the 5G ecosystem. The lack of end-2-end (E2E) testbeds or digital twins for AI-driven private 5G networks is detrimental to the development of innovative and sustainable 5G-IoT services and business models.

6. Conclusion

This article presented a conceptual business model framework for AI-based private 5G-IoT networks. The conceptual framework, formulated from a strategic and technological perspective, consists of ten inter-connected components that can be adopted by private network operators. The technical aspects and associated applications of AI-based private 5G IoT networks were highlighted. In addition, a discussion of the industrial applications of the proposed conceptual framework was presented. In particular, typical business models, vertical industry target markets, key drivers and exemplary use cases were discussed. Finally, critical aspects of a sustainable business model that must be taken into account when implementing AI-based private 5G-IoT networks were highlighted.

Since the current work is based on a theoretically formulated conceptual framework, future research will be implemented on an E2E testbed with machine learning-based analysis applied to the collected primary data to support the development of sustainable business models and innovative services for current and future 6G AI/IoT-based smart and sustainable industries and the preservation of natural ecosystems.

References

[1] J. Whalley, P. Curwen, "Creating value from 5G: The challenge for mobile operators," *Telecommunications Policy*, vol. 48, no. 2, 2024, doi:10.1016/j.telpol.2023.102647.

[2] E. Esenogho, K. Djouani, A.M. Kurien, Integrating Artificial Intelligence Internet of Things and 5G for Next-Generation Smartgrid: A Survey of Trends Challenges and Prospect, *IEEE Access*, vol. 10, , 2022, doi:10.1109/ACCESS.2022.3140595.

[3] A. Ghasempour, Internet of things in smart grid: Architecture, applications, services, key technologies, and challenges, *Inventions*, vol. 4, no. 1, 2019, doi:10.3390/inventions4010022.

[4] D.C. Nguyen, M. Ding, P.N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, H.V. Poor, "6G Internet of Things: A Comprehensive Survey," *IEEE Internet of Things Journal*, vol. 9, no. 1, 2022, doi:10.1109/JIOT.2021.3103320.

[5] J. Tan, X. Sha, T. Lu, B. Dai, "A Short Survey on Future Research of AI and IoT Technologies," in *2022 International Wireless Communications and Mobile Computing, IWCMC 2022*, 2022, doi:10.1109/IWCMC55113.2022.9825425.

[6] A. Yarali, Intelligent Connectivity: AI, IoT, and 5G, 2021, doi:10.1002/9781119685265.

[7] M. Attaran, "The impact of 5G on the evolution of intelligent automation and industry digitization," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 5, 2023, doi:10.1007/s12652-020-02521-x.

[8] P. Curwen, J. Whalley, "Private Networks: A regular column on the information industries," *Digital Policy, Regulation and Governance* , vol. 23, no. 4, 2021, doi:10.1108/DPRG-06-2021-181.

[9] L. Banda, M. Mzyece, F. Mekuria, "5G Business Models for Mobile Network Operators - A Survey," *IEEE Access*, vol. 10, , 2022, doi:10.1109/ACCESS.2022.3205011.

[10] L. Banda, M. Mzyece, F. Mekuria, "Business Models for 5G-IoT Private Networks," in *IEEE AFRICON Conference*, 2023, doi:10.1109/AFRICON55910.2023.10293454.

[11] H. Frank, C. Colman-Meixner, K.D.R. Assis, S. Yan, D. Simeonidou, "Techno-Economic Analysis of 5G Non-Public Network Architectures," *IEEE Access*, vol. 10, , 2022, doi:10.1109/ACCESS.2022.3187727.

[12] S. Mishra, A.R. Tripathi, "IoT Platform Business Model for Innovative Management Systems," *International Journal of Financial Engineering*, vol. 07, no. 03, 2020, doi:10.1142/s2424786320500309.

[13] L. Banda, M. Mzyece, F. Mekuria, "Business Models for 5G and Future Mobile Network Operators," in *Proceedings - 2022 IEEE Future Networks World Forum, FNWF 2022*, 2022, doi:10.1109/FNWF55208.2022.00045.

[14] M. Moussaoui, E. Bertin, N. Crespi, "Divide and Conquer: A Business Model Agenda for Beyond-5G and 6G," *IEEE Communications Magazine*, vol. 61, no. 7, 2023, doi:10.1109/MCOM.001.2200748.

[15] K. Shafique, B.A. Khawaja, F. Sabir, S. Qazi, M. Mustaqim, Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT Scenarios, *IEEE Access*, vol. 8, , 2020, doi:10.1109/ACCESS.2020.2970118.

[16] X. Foukas, G. Patounas, A. Elmokashfi, M.K. Marina, Network Slicing in 5G: Survey and Challenges, *IEEE Communications Magazine*, vol. 55, no. 5, 2017, doi:10.1109/MCOM.2017.1600951.

[17] M. Wen, Q. Li, K.J. Kim, D. Lopez-Perez, O. Dobre, H.V. Poor, P. Popovski, T. Tsiftsis, "Private 5G Networks: Concepts, Architectures, and Research Landscape," *IEEE Journal on Selected Topics in Signal Processing*, vol. 16, no. 1, 2022, doi:10.1109/JSTSP.2021.3137669.

[18] B.S. Khan, S. Jangsher, A. Ahmed, A. Al-Dweik, "URLLC and eMBB in 5G Industrial IoT: A Survey," *IEEE Open Journal of the Communications Society*, vol. 3, , 2022, doi:10.1109/OJCOMS.2022.3189013.

[19] C. Zott, R. Amit, L. Massa, The business model: Recent developments and future research, *Journal of Management*, vol. 37, no. 4, 2011, doi:10.1177/0149206311406265.

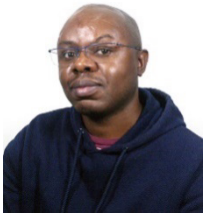
[20] A. Osterwalder, Y. Pigneur, "Investigating the use of the business model concept through interviews," *Shaping Business Strategy in a Networked World, Vols 1 and 2, Proceedings*, 2004.

[21] S. Leshem, V. Trafford, Overlooking the conceptual framework, *Innovations in Education and Teaching International*, vol. 44, no. 1, 2007, doi:10.1080/14703290601081407.

[22] H. Han, J. Yao, Y. Wu, Y. Dou, J. Fu, "Quantum communication based cyber security analysis using artificial intelligence with IoMT," *Optical and Quantum Electronics*, vol. 56, no. 4, 2024, doi:10.1007/s11082-023-06185-7.

[23] M. Eshaghi, R. Rashidzadeh, "An Energy Harvesting Solution for IoT Devices in 5G Networks," in *Canadian Conference on Electrical and Computer Engineering*, 2020, doi:10.1109/CCECE47787.2020.9255802.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



LAURENCE BANDA received the BEng degree in Electrical and Electronics Engineering from the University of Zambia, Lusaka, Zambia in 2006, the MSc in Electronics and Communications Engineering from EISEE, Paris, France in 2011, the MEng in Electrical Engineering from the Tshwane University of Technology, Pretoria, South Africa in 2012 and the PhD in Technology and Operations Management from the University of the Witwatersrand, Johannesburg, South Africa in 2024.

He worked for Huawei Technologies in South Africa for 10 years where he held various positions including Project Manager for Wireless Energy Management, Wireless Advisory Engineer and Technical Trainer. He has published widely in the areas of business models for wireless networks. His current research focus is on technology and innovation management for next-generation wireless networks incorporating AI and IoT.

Received: 11 September, 2024, Revised: 12 October, 2024, Accepted: 12 October, 2024, Online: 22 October, 2024

DOI: https://dx.doi.org/10.55708/js0310003

Navigating the Autonomous Era: A Detailed Survey of Driverless Cars

Vaibhavi Tiwari*

Vaibhavi Tiwari, School of Computing, Montclair State University, New Jersey, USA

*Corresponding author: Vaibhavi Tiwari, Montclair State University, Email: tiwariv1@montclair.edu

ABSTRACT: The incorporation of cutting-edge technologies like sensor networks, artificial intelligence (AI), and vehicle-to-everything (V2X) communication has hastened the rollout of autonomous vehicles (AVs), offering significant possibilities for the future of transportation. This document offers an extensive overview of AV technology, covering essential elements such as technological infrastructure, degrees of automation, cybersecurity threats, societal impacts, regulatory structures, and emerging trends. This analysis emphasizes the existing obstacles and progress within the industry by examining the activities of key entities like Tesla, Waymo, and General Motors. Additionally, a comparative examination of autonomous vehicles and drones is performed, providing distinct perspectives on possible cybersecurity vulnerabilities shared by both technologies, including GPS spoofing, jamming, and unauthorized data interception. This multifaceted approach highlights not only the existing vulnerabilities but also proposes proactive measures that can be implemented to reduce comparable risks across various AV platforms. The results highlight the necessity of establishing strong cybersecurity measures, overcoming regulatory challenges, and building public confidence to realize the complete promise of autonomous vehicles as secure, effective, and eco-friendly transportation options. This analysis provides an essential resource for comprehending the complex aspects of AV technology and its consequences, offering readers a comprehensive perspective on the challenges and opportunities within the autonomous vehicle sector.

KEYWORDS: Driverless cars, autonomous vehicles, sensors, AI, V2X communication, SAE levels of automation, cybersecurity

1. Introduction

The automotive sector is swiftly progressing towards a future characterized by autonomous vehicles (AVs), with driverless cars at the forefront of this evolution. The vehicles utilize state-of-the-art technologies, such as intricate sensor networks, advanced artificial intelligence (AI), and real-time data processing, which together hold the potential to transform transportation by improving safety, efficiency, and convenience. Nonetheless, the incorporation of these complex systems brings forth significant cybersecurity issues that need to be tackled to safeguard AV performance and guarantee user safety.

Historically, autonomous vehicle research has spanned several decades, marking the transition of AVs from speculative fiction to a reality on our roads today. Experiments in this field began as early as the 1920s with Ralph Teetor's invention of cruise control, and later expanded into semi-autonomous systems developed by Japan's Tsukuba Mechanical Engineering Laboratory in the 1970s. These early projects laid the foundation for subsequent advancements, including Carnegie Mellon University's Navlab and ALV projects, which achieved milestones in autonomous cross-country travel by the mid-1990s [1, 2, 3]. Government support, such as the United States' $650 million allocation for the National Automated Highway System in the 1990s [4], further bolstered AV innovation, eventually enabling private industry players like Waymo, Tesla, and Nuro to lead commercial deployments from the late 2010s onward [5, 6].

In recent years, the concept of AVs has rapidly advanced due to technological progress, transitioning from controlled test environments to limited public road usage. Notably, McKinsey's 2023 global executive survey highlights significant advancements, with insights from 86 decision-makers forecasting the commercial availability of Level 4 (L4) autonomous vehicles and robo-taxis by 2030. Despite these advancements, substantial financial investments remain necessary. For instance, developing fully autonomous trucks and L4/L5 robo-taxis requires over $4 billion and $5 billion, respectively [7]. This ongoing progress underscores the industry's commitment but also highlights the resource-intensive nature of AV development.

This study uniquely contributes to the AV landscape by conducting a comparative analysis between AVs and drones. This comparison is particularly valuable because drones have encountered a wide array of cybersecurity threats. By examining these threats, the research explores how similar vulnerabilities could manifest within AV systems, thereby offering a basis for understanding potential misuse of AV technology. This insight allows for proactive mitigation strategies that can be applied to both AVs and drones, enhancing our overall preparedness against cyber threats.

However, the adoption of AVs at scale faces numerous hurdles, including regulatory, safety, and cybersecurity concerns. The cybersecurity landscape is especially complex, with threats such as remote exploits, unauthorized access, and adversarial machine learning attacks posing significant risks to AV systems [8]. The increasing reliance on wireless communication in AVs exposes them to attacks that could compromise vehicle control, sensitive data, and communication networks. The rapid evolution of cybersecurity risks between 2023 and 2024, such as adversarial attacks against AI systems and supply chain vulnerabilities, further emphasizes the need for robust cybersecurity protocols.

Beyond cybersecurity, AVs must also navigate regulatory and societal challenges. As the industry grows—projected to reach $556.67 billion by 2026—collaboration among automotive sectors, regulatory bodies, and researchers becomes essential to address the socioeconomic impacts, such as job displacement within driving-related industries. Trust-building initiatives, such as Waymo's extensive road testing and Tesla's Full Self-Driving (FSD) program, continue to play a critical role in driving public acceptance [6, 9, 10].

Through this study, readers will gain insights into AV technologies, the associated cybersecurity risks, and the broader societal and regulatory challenges. By presenting a detailed comparative analysis with drones, this research not only identifies shared vulnerabilities but also provides a framework for preemptive mitigation strategies. This exploration is intended to equip readers with a holistic understanding of AV technology, guiding them through the steps needed for safe, efficient, and socially responsible autonomous transportation. The projected growth trends [11] for the autonomous vehicle market are illustrated in Figure 1.



Figure 1: Statistics of the Autonomous Vehicle Market (2023 - 2033)

2. Motivation

There are a number of powerful incentives that are driving the development of autonomous vehicles. Each of these incentives addresses significant societal, economic, and environmental challenges respectively:

2.1. Enhancing Road Safety

An important driving force behind the development of autonomous vehicles is the ability to improve road safety by greatly reducing human error, which is a major contributor to traffic accidents. As per the World Health Organization, some 1.3 million individuals perish annually in road traffic accidents, with the bulk of these incidents being caused by human error [12, 13]. Driverless cars utilize sophisticated sensors, machine learning algorithms, and real-time data processing to enhance decision-making capabilities, surpassing those of human drivers. This results in a decreased probability of accidents and a significant preservation of human life [14].

2.2. Increasing Transportation Efficiency

Through the optimization of vehicle movement and the reduction of the amount of time spent driving, autonomous vehicles have the potential to change the efficiency of transportation. It is possible for autonomous vehicles to establish contact with one another as well as with traffic management systems. This allows them to coordinate their movements, limit the number of occasions in which they are forced to stop and go, and determine the routes that are ultimately the most efficient. It is possible that these outcomes will result in greater traffic efficiency, decreased traffic congestion, and shorter travel durations [15, 16]. These outcomes are beneficial to individual commuters as well as the economy as a whole because they will increase productivity and decrease fuel consumption simultaneously [17, 18].

2.3. Reducing Traffic Congestion

Traffic congestion has a substantial impact on urban areas, leading to inefficiency in terms of time management, increased levels of pollution, and economic setbacks. Through the synchronization of their movements, the optimization of the timing of traffic lights, and the reduction of the need for parking spots in densely populated metropolitan areas, autonomous cars have the potential to alleviate traffic congestion. The integration of vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication gives autonomous cars the capacity to dynamically adjust their speed and routes in order to avoid traffic congestion and maintain a continuous movement of vehicles. Such a feature allows autonomous vehicles to maintain a consistent movement of vehicles [19, 20, 21].

2.4. Improving Mobility for All

Autonomous vehicles have the potential to significantly improve the transportation alternatives available to individuals who are unable to operate a vehicle, such as the elderly, the disabled, or those who do not have their own means of transportation [22]. Individuals who may have difficulties with mobility can benefit from the on-demand transportation services provided by autonomous vehicles, which offer a reliable and convenient mode of transportation. This has the potential to improve self-sufficiency, facilitate access

to essential services, and eventually lead to an improvement in the overall quality of life for the aforementioned demographics [23].

2.5. Contributing to Environmental Sustainability

The optimization of driving patterns, the decrease of idle time, and the encouragement of the use of electric vehicles (EVs) are all ways in which autonomous vehicles have the potential to improve environmental sustainability. Because they are able to operate more efficiently and eliminate excessive acceleration and braking, driverless cars have the potential to reduce the amount of fuel that is consumed as well as the emissions of greenhouse gases [24]. A large number of projects involving autonomous vehicles are currently being developed concurrently with the development of electric vehicle technology. This has the potential to significantly reduce the negative impact that transportation has on the environment [25]. The incorporation of autonomous technology into electric vehicles (EVs) has the potential to hasten the adoption of clean energy within the transportation sector, thereby making a contribution to broader environmental goals [24, 26].

2.6. Economic Benefits

There is a high probability that the introduction of autonomous vehicles will result in considerable economic benefits. It is possible to realize significant cost savings in the areas of healthcare, emergency services, and vehicle repairs by reducing the number of accidents that occur on the roads. A large reduction in the number of accidents that occur in the United Kingdom, for instance, might result in savings of almost two billion pounds by the year 2030 at the very least [27]. Furthermore, it is predicted that the autonomous automobile industry will provide a multiplicity of employment opportunities in the areas of technological innovation, infrastructure advancements, and mobility services. This sector would also deliver more than £51 billion in economic benefits [28].

2.7. Accessibility and Inclusion

Autonomous vehicles have the potential to provide individuals with disabilities, the elderly, and other individuals who do not drive with enhanced mobility options, thereby boosting their freedom and general well-being [22]. Autonomous cars have the potential to be designed in such a way that they can accommodate a wide range of accessibility requirements, thereby offering a reliable and safe mode of transportation to all users.

3. Technologies Behind Driverless Cars

The technology used in driverless automobiles is varied and intricate, encompassing the integration of advanced sensors, artificial intelligence, vehicle-to-vehicle and vehicle-to-infrastructure communication, and sophisticated decision-making algorithms. These technologies collectively allow autonomous vehicles to accurately detect their surroundings, precisely determine their location, strategize safe routes, and efficiently navigate while interacting with other vehicles and infrastructure. Table 1 provides a brief overview of these technologies, highlighting the essential components used in each.

Table 1: An Overview of Technologies

Technology	Description
Sensors and Perception Systems	Uses LiDAR, Radar, Cameras, and Ultrasonic sensors for environmental mapping. Provides real-time detection of surroundings.
Localization and Mapping	GPS, IMUs, and SLAM for positioning. Enables navigation in complex environments.
Decision-Making Algorithms	Path planning, obstacle avoidance, predictive modeling. Adapts to real-time traffic conditions.
Vehicle-to-Vehicle (V2V) Communication	Shares data on speed, position, and intentions. Improves coordination and safety.
Vehicle-to-Infrastructure (V2I) Communication	Interacts with infrastructure like traffic lights. Optimizes traffic flow.
Artificial Intelligence and Machine Learning	Deep learning for perception and decision-making. Reinforcement learning for pattern recognition.

3.1. Sensors and Perception Systems

The powerful sensors and vision systems used in autonomous vehicles are central to their operational capabilities. These systems combine various types of sensors to collect and interpret data from the surrounding environment, generating a comprehensive view. LiDAR (Light Detection and Ranging) systems emit laser pulses, measuring the time it takes for them to reflect off objects, which helps create high-resolution 3D maps [29]. This is crucial for identifying the shape, distance, and size of obstacles, enabling precise environmental mapping and navigation. Radar sensors, which use radio waves to detect the distance and speed of objects, are essential for adaptive cruise control and collision avoidance, and are particularly effective in adverse weather where optical sensors may fail [30]. Visual cameras capture images of the environment that are processed using computer vision algorithms for tasks such as object recognition, lane detection, and traffic sign identification. Modern autonomous vehicles typically deploy multiple cameras to cover various angles, creating a complete visual representation of the surroundings. Ultrasonic sensors [31], primarily used for short-range detection, assist in parking maneuvers by detecting objects close to the vehicle, ensuring safety during low-speed operations. Collectively, these perception systems synthesize data from multiple sources to produce an accurate model of the vehicle's surroundings, which is fundamental for decision-making and control activities.

3.2. Localization and Mapping

Accurate localization and mapping are essential for autonomous vehicle navigation, as they enable vehicles to determine their precise position and navigate complex environments. The Global Positioning System (GPS) provides basic positioning information, but in urban settings, where signal blockages by tall buildings can occur, GPS is often supplemented by other localization technologies. Inertial Measurement Units (IMUs) track the vehicle's orientation and motion, supplying critical data for dead-reckoning

and enhancing GPS accuracy. Simultaneous Localization and Mapping (SLAM) is another key technology, involving algorithms that build a map of the environment while simultaneously tracking the vehicle's position within it. SLAM integrates data from LiDAR, cameras, and IMUs to produce detailed and accurate maps, which are vital for navigating dynamic and unfamiliar surroundings. Advanced navigation systems often employ a hybrid approach, combining fundamental maps with real-time perception data to adapt to immediate environmental changes [32]. Innovations from institutions like MIT's CSAIL have even developed systems that rely on sparse topological maps and real-time sensor data, allowing autonomous vehicles to navigate without detailed maps.

3.3. Decision-Making Algorithms

Autonomous vehicles rely on sophisticated decision-making algorithms to interpret sensor data and execute driving tasks safely and effectively. Path planning algorithms determine the optimal route from the vehicle's current position to its destination, using techniques such as graph-based search (e.g., A* algorithm) and optimization methods to ensure collision-free paths. These algorithms also dynamically adjust to real-time traffic conditions and obstacles [33]. Obstacle avoidance algorithms detect and navigate around obstacles with inputs from LiDAR, radar, and cameras, utilizing methods like Detection and Tracking of Moving Objects (DATMO) to forecast obstacle movements and adjust the vehicle's path accordingly [33]. Additionally, predictive modeling anticipates the actions of other road users, such as pedestrians or other vehicles, by applying machine learning models, including deep neural networks, to enhance the accuracy of these predictions. Together, these decision-making processes enable the vehicle's control system to analyze the environment, adhere to traffic rules, and ensure secure and efficient movement.

3.4. Vehicle-to-Vehicle (V2V) Communication

Vehicle-to-Vehicle (V2V) communication enables autonomous cars to share real-time information on their speed, position, and driving intentions. This exchange of data improves situational awareness and facilitates coordinated maneuvers, such as platooning, where vehicles travel in close formations to reduce aerodynamic drag and enhance fuel efficiency. V2V communication ensures precise coordination of speed and braking among platooned vehicles, which bolsters both safety and efficiency. Furthermore, V2V is crucial for accident prevention, as it allows vehicles to relay alerts about potential hazards, sudden stops, or other critical incidents [34].

3.5. Vehicle-to-Infrastructure (V2I) Communication

Vehicle-to-Infrastructure (V2I) communication involves interactions between vehicles and road infrastructure, such as traffic lights, road signs, and other smart systems. This technology plays a key role in optimizing traffic flow, as vehicles can receive real-time updates on traffic conditions, signal timings, and available detours, thereby reducing congestion and improving travel times. V2I communication also enhances road safety by providing vehicles with warnings about hazards, construction zones, or changes in road conditions. Integration with smart city infrastructure further extends V2I's capabilities, enabling comprehensive traffic control and significant improvements in urban transportation [34].

3.6. Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) and Machine Learning (ML) are critical components of the architectures that enable autonomous driving. These technologies empower vehicles to learn from data, recognize patterns, and make decisions in complex situations. Deep Learning, for instance, is used extensively for visual perception, object recognition, and decision-making. Through deep neural networks, autonomous vehicles can process inputs from various sensors to detect objects, recognize traffic signs, and predict the actions of other road users. Reinforcement Learning, on the other hand, allows vehicles to learn from real-world driving experiences and continually enhance their performance. This adaptive approach enables autonomous systems to refine their decision-making processes over time and respond effectively to new environments. Overall, AI systems in autonomous vehicles facilitate the interpretation of sensory data, enable situational awareness, and support safe navigation without human intervention, improving their capabilities as they encounter diverse driving conditions.

4. Levels of Autonomy

4.1. SAE Levels of Automation

The Society of Automotive Engineers (SAE) has defined six levels of vehicle automation, from Level 0 (no automation) to Level 5 (full automation), which describe the degree of driver intervention required and the vehicle's capabilities at each stage [35]. These levels provide a framework to understand the evolution and technical capabilities of AV systems:

1. **Level 0 (No Automation)**: At this level, the human driver is fully responsible for controlling the vehicle. While some technologies, such as warnings or momentary assistance (e.g., collision alerts), may be present, they do not control the vehicle.

2. **Level 1 (Driver Assistance)**: Basic driver-assist technologies, such as adaptive cruise control or lane-keeping assistance, enable limited control over either steering or acceleration/deceleration, but not simultaneously. Current technologies at this level often employ sensors and basic AI to interpret lane markings or maintain a safe following distance.

3. **Level 2 (Partial Automation)**: The vehicle can control both steering and speed under certain conditions, using a combination of radar, LiDAR, and camera systems to monitor the environment. Although the driver remains responsible for monitoring the road and staying alert, this level marks the transition to shared control.

4. **Level 3 (Conditional Automation)**: Vehicles at this level can manage all driving tasks under specific conditions, such as highway driving. However, the driver must be ready to take over when the system requests. The transition between automated and manual control relies on advanced sensor fusion and environmental mapping, allowing the system to make decisions based on real-time data processing.
5. **Level 4 (High Automation)**: At this stage, the vehicle can perform all driving functions and monitor the environment in designated operational design domains (ODDs), such as urban areas or specific weather conditions, without driver intervention. While human oversight is not needed within the ODD, technological limitations prevent full autonomy under all conditions, with industry leaders like Waymo and Cruise currently testing Level 4 systems in pilot.
6. **Level 5 (Full Automation)**: This level represents the goal of fully autonomous vehicles, capable of performing all driving tasks in all conditions without any human intervention. The technical challenges include creating a robust infrastructure of AI, machine learning, and V2X communication, but full industry adoption remains unrealized as of 2024.

Industry Adoption and Comparison: While significant strides have been made in Levels 2 and 3, full Level 5 automation is not yet achieved. Currently, automakers and tech firms, such as Tesla and Waymo, are focusing on improving Levels 3 and 4 with varied strategies. Tesla's approach, for example, relies heavily on computer vision and advanced neural networks for higher levels of autonomy, while Waymo incorporates high-definition mapping and extensive LiDAR systems. The comparative analysis within this paper explores these strategies, emphasizing how different companies prioritize elements like sensor fusion and machine learning to overcome the unique challenges at each automation level. By highlighting these approaches, the paper outlines key industry practices that contribute to the broader landscape of AV technology development [35].

4.2. Gradual Progression Towards Full Autonomy

The transition towards complete autonomy is slow, with cars now functioning at either Level 2 or Level 3. These levels enable vehicles to do specific driving tasks while still necessitating human supervision. The transition to higher levels of automation will be driven by notable progress in technology and legal frameworks. For instance, the testing and implementation of Level 3 systems such as Honda's "Traffic Jam Pilot" and Mercedes-Benz's Drive Pilot demonstrate encouraging advancements in minimizing human intervention in particular situations.

Autonomous vehicles must address various challenges before achieving full autonomy, including:

- **Advanced Software and Mapping**: Ensuring the software can handle diverse driving conditions and environments.
- **Human Factors**: Developing systems that can safely transfer control between human drivers and autonomous systems.
- **Regulatory and Legal Frameworks**: Establishing standards for liability, safety, and operation of autonomous vehicles.
- **Ethical and Security Concerns**: Addressing ethical dilemmas such as the "trolley problem" and ensuring robust cybersecurity measures.

The continued research and development efforts that are being made by industry leaders, government agencies, and academic institutions are absolutely necessary in order to overcome these hurdles and make progress toward completely autonomous vehicles.

5. Current State-of-the-Art

5.1. Industry Leaders and Major Players

Several important players who are at the forefront of developing and deploying autonomous vehicle (AV) technology are present in the industry of driverless cars, which is characterized by the presence of these key players. The following are notable businesses:

Tesla: Tesla has made great progress with its Full Self-Driving (FSD) software, which promises to provide full autonomy through continual over-the-air upgrades. This software has undergone significant development. Rather than relying on LiDAR, Tesla's strategy makes use of a vision-based system that is equipped with cameras. Additionally, the company extensively depends on artificial intelligence and neural networks to interpret visual data.

Waymo: Waymo, which happens to be a part of Alphabet Inc., is widely regarded as a pioneer in the field of autonomous vehicles. Waymo's autonomous driving systems are able to reach a high level of precision and reliability thanks to the utilization of a combination of LiDAR, radar, and cameras. Waymo's self-driving taxis, which are currently operating in Phoenix, Arizona, have delivered thousands of rides, in addition to providing critical data from the real world.

Uber: The Advanced Technologies Group (ATG) of Uber has been contributing to the development of autonomous vehicles (AVs) for its ride-hailing services. This group has been working on self-driving technologies. In order to enhance the safety and effectiveness of its autonomous systems, Uber has carried out extensive testing in a variety of metropolitan situations.

General Motors (GM) and Ford: GM and Ford, established car manufacturers, are making significant financial commitments to autonomous technology through their respective companies, Cruise and Argo AI. These firms are investigating several uses of autonomous vehicles (AVs), such as ride-sharing and delivery services, by utilizing their substantial knowledge and infrastructure in manufacturing.

5.2. Commercial Deployments and Pilot Programs

When it comes to the collection of data and the improvement of autonomous vehicle technology, commercial deployments and pilot programs are absolutely necessary. There are many noteworthy programs, including:

Waymo One: Waymo's autonomous taxi service has been operational in Phoenix, Arizona, since December 2018,

and recently expanded to San Francisco and Los Angeles. Within the confines of a geofenced area, this business offers trips that are completely autonomous to the general public, collecting vital data that may be used to improve their technology.

Tesla Full Self-Driving (FSD) Beta: Thousands of Tesla owners are participating in the FSD beta program, which is evaluating the software under real-world settings while it is being developed. Continuous feedback is provided to Tesla by this program, which enables incremental changes to be made to the capabilities of the FSD capability.

Cruise by GM: Cruise has been doing extensive testing of its driverless vehicles in San Francisco, with a particular emphasis on testing them in urban situations. In order to capitalize on General Motors' manufacturing skills, Cruise intends to establish a commercial robotaxi service shortly.

Argo AI and Ford: A number of cities, including Miami and Austin, are participating in pilot programs that are being carried out by Argo AI in conjunction with Ford. Within the context of ride-sharing and delivery services, these activities are intended to improve the autonomous driving systems in preparation for their eventual deployment in commercial settings.

5.3. Technological Limitations and Challenges

In spite of the progress that has been made, autonomous cars continue to confront a number of technological restrictions and opportunities:

Complex Urban Environments: This is a challenge for autonomous vehicles because of the unpredictability and complexity of urban environments. The use of autonomous vehicles (AV) presents considerable issues in situations where there is a high volume of traffic, pedestrians, cyclists, and emergency vehicles.

Adverse Weather Conditions: The reliability of autonomous vehicles (AV) systems can be negatively impacted by weather conditions such as rain, snow, and fog, which can hinder the functioning of sensors like as cameras and LiDAR. The creation of reliable systems that are capable of functioning successfully in any and all weather circumstances continues to be a significant problem.

Cybersecurity: When it comes to protecting autonomous vehicles (AVs) against hacking and data breaches, it is necessary to provide comprehensive cybersecurity safeguards. Because of the high level of connectivity that exists amongst autonomous vehicles, any vulnerabilities that may exist could be exploited, which could result in the theft of important data or the malicious control of the vehicle.

Regulatory and Ethical Issues: A thorough regulatory framework must be developed in order to facilitate the broad use of autonomous vehicles (AVs). The establishment of standards for safety, liability, and insurance is included in this requirement. In addition, it is necessary to handle ethical conundrums, such as the process of making decisions in situations where accidents are unavoidable.

To summarize, although there have been notable progressions, it is imperative to do further research and development in order to surmount these obstacles and fully realize the capabilities of autonomous cars.

6. Existing Cybersecurity Threats

Driverless cars are equipped with a multitude of interconnected systems that manage everything from navigation to communication with external devices. These systems are susceptible to various types of cyberattacks, including:

6.1. Remote Exploits and Unauthorized Access

The dependence on wireless connectivity renders autonomous vehicles susceptible to remote exploitation. Unauthorised gain of access to the vehicle's control systems by attackers has the potential to result in catastrophic failures. As an illustration, in 2015, security researchers Charlie Miller and Chris Valasek successfully conducted a remote hack on a Jeep Cherokee by taking advantage of weaknesses in its Uconnect smart entertainment system. Unauthorized access to the vehicle's internal network was successfully obtained, enabling the delivery of commands to vital systems such as the engine, gearbox, and brakes. The cyber assault demonstrated that vehicles with unsecured connectivity might be influenced from any location, therefore giving rise to significant apprehensions regarding the security of autonomous vehicles [36].

6.2. Data Breaches and Privacy Concerns

Autonomous vehicles amass huge quantities of data, encompassing personal information as well as driving trends. Data breaches have the potential to result in the unauthorised acquisition and improper use of this confidential data. The 2016 Uber data breach, which compromised the personal data of 57 million passengers and drivers, highlights the possible hazards linked to the extensive data retention inherent in autonomous vehicles. Uber's former Chief Security Officer, Joe Sullivan, endeavored to conceal the breach by offering the hackers $100,000 as part of a "bug bounty" scheme, resulting in severe legal repercussions [37].

6.3. Vehicle-to-Everything (V2X) Communication Attacks

Driverless cars communicate with infrastructure, other vehicles, and even pedestrians through Vehicle-to-Everything (V2X) technology. These communications are critical for safe operation but are vulnerable to various types of attacks.

In 2019, researchers demonstrated the feasibility of spoofing GPS signals to mislead autonomous vehicles, potentially causing them to veer off course or crash. This attack involves transmitting fake GPS signals that override the vehicle's legitimate signals, leading to dangerous miscalculations [38].

V2X communication also relies on wireless signals that can be jammed, disrupting the flow of crucial information. In 2021, a jamming attack was demonstrated where multiple vehicles were rendered unable to receive or send data, causing significant traffic disruptions and even roadblocks. These incidents highlight the vulnerabilities in V2X communication protocols that could be exploited to create gridlock or force vehicles into unsafe conditions [39].

Additionally, relay attacks have been used to intercept and delay V2X messages, causing vehicles to react to out-

dated or incorrect information. This can lead to dangerous situations where vehicles may fail to stop at red lights or collide due to delayed braking signals [40].

6.4. Sensor Spoofing Attacks

Vehicle perception systems, including LiDAR, radar, and cameras, are vulnerable to sensor spoofing. Attackers have the ability to create deceptive signals that can be mistaken for genuine objects by these sensors. As an illustration, studies have shown that by projecting certain patterns onto road signs, the vehicle's image recognition system can be tricked into misinterpreting the sign, which could result in risky decisions. In a similar vein, through the use of laser spoofing, malicious individuals have the ability to manipulate LiDAR readings. This can result in the creation of deceptive obstacles or the concealment of genuine ones [41].

6.5. Malware and Software Exploits

Vehicle software can be targeted by malware attacks that take advantage of weaknesses in the operating system or third-party applications. As an example, in 2018, a group of experts made an important finding regarding a vulnerability in Tesla's Model S. This finding enabled them to remotely control the vehicle. The attack was carried out by taking advantage of a vulnerability in the vehicle's browser, which enabled the researchers to insert harmful code. Once inside, they had the ability to control various aspects of the car, including unlocking doors and even steering [42].

6.6. Man-in-the-Middle (MitM) Attacks

Intercepting and potentially altering communications between the vehicle and external systems is a common occurrence in MitM attacks. In 2020, a groundbreaking discovery was made when experts successfully showcased a method of intercepting V2X communication, resulting in a potential vulnerability for Tesla vehicles. The manipulation of sensor data led to the vehicle misinterpreting its surroundings. Such an attack has the potential to result in the vehicle receiving inaccurate information, such as incorrect speed limits or traffic signals, which can lead to unsafe driving behavior [43].

6.7. Denial-of-Service (DoS) Attacks

Denial of Service (DoS) attacks aim to incapacitate a vehicle's systems or communication channels by inundating them with excessive traffic. In a recent experiment conducted in 2019, a group of researchers successfully demonstrated a disruptive attack on a fleet of autonomous vehicles. The attack involved overwhelming the vehicles' communication channels with an excessive amount of traffic, causing them to experience a denial-of-service (DoS) situation. This incident resulted in a total disruption of vehicle-to-vehicle (V2V) communication, resulting in a breakdown of synchronized driving and the possibility of accidents [44].

7. Emerging Threats in 2023-2024

Recent advancements in AI and machine learning have introduced new dimensions to cybersecurity threats in driverless cars. The following are some of the most concerning threats identified in the past year:

7.1. Adversarial Machine Learning (AML) Attacks

As driverless cars increasingly rely on AI to make real-time decisions, they become targets for adversarial machine learning (AML) attacks. These attacks exploit vulnerabilities in AI models by introducing carefully crafted perturbations to the input data. These perturbations, often imperceptible to human observers, can cause the AI to make incorrect decisions, such as misclassifying road signs or obstacles. For instance, in 2023, researchers demonstrated that by placing inconspicuous stickers on stop signs, they could cause AI systems in driverless cars to misinterpret them as yield signs, leading to potentially dangerous outcomes. The technical complexity of AML attacks lies in the ability to generate perturbations that evade detection while consistently fooling the model across various scenarios, making them a significant threat to the safety and reliability of autonomous vehicles [45].

7.2. Supply Chain Attacks

The global supply chain for automotive components is vast and interconnected, making it a prime target for sophisticated cyberattacks. In 2024, a major automotive supplier was compromised through a multi-stage attack that involved the insertion of malware into firmware updates for critical components used in driverless cars. The malware was designed to remain dormant, undetected by traditional security measures, until specific conditions were met, at which point it would activate and allow remote attackers to take control of the affected vehicles. This attack exemplifies the growing complexity and reach of supply chain threats, where the integrity of software and hardware components can be compromised at any point along the supply chain, from manufacturing to distribution. The challenge in defending against these attacks lies in the need for comprehensive verification processes and the integration of secure development practices across all tiers of the supply chain [46].

7.3. Ransomware Targeting Vehicle Systems

Ransomware targeting AI-driven and autonomous vehicles represents an emerging cybersecurity challenge due to the complexity and connectivity of these systems. Such attacks could exploit vulnerabilities in the vehicle's software stack, potentially locking down critical systems like navigation, braking, or powertrain control. By encrypting the vehicle's control systems, ransomware [47] can render the car inoperable until a ransom is paid, usually in cryptocurrency. The decentralized nature of these attacks makes them difficult to trace, and the integration of AI increases the risk, as machine learning models can be manipulated or disrupted to exacerbate the impact. Additionally, the connected ecosystem of vehicles, with constant communication to cloud services and other devices, introduces multiple attack vectors, such as over-the-air updates, that can be compromised. This highlights the urgent need for multi-layered

security strategies, including robust encryption, anomaly detection systems, and secure software development practices to mitigate these risks [48].

7.4. Keyless Car Theft and Relay Attacks

The advent of key less entry and ignition systems in modern vehicles has introduced new security vulnerabilities in the automotive industry. These conveniences have become targets for sophisticated theft techniques, particularly man-in-the-middle attacks. Such attacks exploit the communication between cars and key fobs, allowing thieves to unlock and start vehicles without triggering alarms [49]. In 2023, the discovery of Bluetooth relay attacks on Tesla vehicles highlighted another critical weakness. This exploit took advantage of vulnerabilities in Bluetooth Low Energy (BLE) protocols used in passive entry systems. By relaying Bluetooth signals, attackers could access vehicles without the physical presence of the key fob. Although Tesla addressed this specific issue with a software update, the incident underscored a broader cybersecurity challenge [50].

7.5. Risks in Autonomous Vehicle Charging Infrastructure

As electric vehicles (EVs) have evolved, cybersecurity vulnerabilities in their charging infrastructure have become a significant concern. These issues are even more critical for autonomous vehicles (AVs), which also rely on secure data exchange with charging stations. The increased complexity and autonomy of AVs heighten the risks, including the potential for malware, fraud, remote manipulation, and disabling of charging stations. These vulnerabilities pose serious threats to the safety and integrity of the autonomous vehicle ecosystem [51].

7.6. Causing Traffic Jams

The interconnected nature of autonomous vehicles makes them vulnerable to coordinated cyberattacks that could lead to significant traffic disruptions. By compromising the control systems of multiple driverless cars simultaneously, an attacker could create artificial traffic jams. For instance, cars could be programmed to slow down or stop in strategic locations, blocking key intersections or highways [52]. Such disruptions could be used as a form of protest, to create chaos in urban environments, or as a precursor to other criminal activities, such as facilitating the escape of criminals from law enforcement by creating diversions.

7.7. Noise Pollution

In recent events, San Francisco residents experienced disturbances due to Waymo's autonomous vehicles persistently honking in residential areas during nighttime hours. This unexpected behavior, initially captured on video by local residents, drew significant public attention and prompted Waymo to respond with an apology and immediate corrective measures. According to Waymo, the honking feature was originally designed to minimize collision risks on public roads by signaling to other vehicles and pedestrians. However, when applied in confined parking areas, the feature led to excessive noise pollution, particularly during early morning hours, causing distress to the community [53]. This incident not only highlights the complexities and unintended consequences of deploying autonomous vehicles in urban environments but also raises concerns about potential misuse. Autonomous vehicles' honking functions could be exploited by cybercriminals to create targeted noise pollution in densely populated areas. If an AV system were to be compromised, attackers could manipulate honking or other sound features, intensifying urban noise pollution and causing widespread disruption. This example underscores the need for robust cybersecurity measures that address not only operational functionality but also safeguard against potential abuses that could impact public well-being. As AV technology continues to advance, these considerations will be crucial for ensuring harmonious integration within urban settings.

8. Countermeasures for AV Cybersecurity Threats

Autonomous vehicles (AVs) encounter various cybersecurity threats that can jeopardize their safety, operational capabilities, and the confidence of the public. This section examines essential strategies aimed at reducing these risks through the improvement of AV system security against a range of potential attacks.

8.1. Adversarial Machine Learning (AML) Attacks

Adversarial Machine Learning (AML) attacks exploit weaknesses in AI models, causing them to make incorrect decisions. To counter AML attacks, AV developers can utilize several advanced defense techniques:

- **Adversarial Training:** This technique involves training AI models on adversarial examples—data intentionally perturbed to mislead the model—so that the model learns to recognize and resist such attacks. By exposing the model to these adversarial inputs during training, its robustness against real-world attacks is improved [54].
- **Defensive Distillation:** This approach reduces the model's sensitivity to adversarial perturbations by smoothing the decision boundaries, making it harder for subtle modifications to alter the AI's output [55].
- **Feature Squeezing:** By reducing the complexity of input data, feature squeezing eliminates extraneous features that could be exploited. Techniques such as image bit depth reduction or smoothing can mitigate adversarial input by stripping away noise and focusing on essential information [56].
- **Robustness Verification:** Formal methods and automated testing frameworks can be applied to verify the resilience of AI models against adversarial perturbations, providing assurances that the models perform reliably under a range of potential attack scenarios [57].

8.2. Supply Chain Attacks

The complexity of AV supply chains introduces vulnerabilities that can be targeted through various means. Addressing these requires both organizational and technical strategies:

- **Blockchain for Component Tracking:** Implementing blockchain can provide transparent and immutable records for each component's origin and journey through the supply chain, helping to ensure that only verified components are integrated into AV systems [58, 59].
- **Secure Boot and Hardware Roots of Trust:** Secure boot mechanisms ensure that AV systems only load authenticated software at startup. Hardware roots of trust further protect the integrity of the system by establishing a secure hardware foundation, preventing malware insertion at the hardware level [58].
- **Code Signing and Verification:** All software updates and components should be cryptographically signed, and systems must verify these signatures before applying any updates. This process ensures that only authorized and untampered code is executed within AV systems.
- **Comprehensive Vulnerability Assessments:** Regular security audits and vulnerability assessments are essential for identifying and mitigating potential supply chain weaknesses. By partnering with trusted third-party cybersecurity firms, manufacturers can gain additional insights into their supply chain's security posture [59].

8.3. Ransomware Attacks Targeting Vehicle Systems

Ransomware poses a growing threat to AV systems by potentially locking critical functionalities. Effective countermeasures include:

- **Strong Encryption and Backup Protocols:** Encrypting key data within the vehicle's system and maintaining secure, regularly updated backups can prevent data loss and facilitate recovery in case of a ransomware attack [47].
- **Endpoint Detection and Response (EDR):** Advanced EDR systems provide real-time monitoring of AV endpoints, detecting unusual patterns that may indicate ransomware activities. Rapid response capabilities enable the containment and neutralization of threats before they escalate.
- **Anomaly Detection Systems:** Employing machine learning models that analyze normal system behaviors can help detect ransomware activity. When abnormal behavior patterns are identified, such as unauthorized encryption attempts, the system can isolate affected modules to prevent further damage [47].

8.4. Keyless Car Theft and Relay Attacks

AVs with keyless entry systems are susceptible to relay attacks that intercept and amplify signals between the key fob and the vehicle. Countermeasures include:

- **Ultra-Wideband (UWB) Technology:** UWB-based systems offer enhanced accuracy in determining the proximity of the key fob to the vehicle, reducing the likelihood of successful relay attacks [60].
- **Two-Factor Authentication (2FA) for Access:** By requiring a secondary authentication step (such as a smartphone verification) [61], AVs can add an additional layer of security that limits unauthorized access.
- **Interference Detection Systems:** These systems monitor and detect signal anomalies associated with relay attacks, enabling AVs to trigger alerts or disable keyless entry temporarily when an attack is suspected.

8.5. Risks in Autonomous Vehicle Charging Infrastructure

AV charging stations are potential targets for cyberattacks that could disrupt vehicle operations. Mitigating these risks involves:

- **Transport Layer Security (TLS) Protocols:** Secure communication between AVs and charging stations through TLS [62] ensures data integrity and confidentiality, making it more challenging for attackers to intercept or manipulate information.
- **Network Segmentation and Firewalls:** Isolating charging stations from broader vehicle and grid networks prevents unauthorized access and contains attacks to specific segments, limiting the scope of potential damage.
- **Regular Cybersecurity Audits:** Continuous monitoring and assessment of charging infrastructure cybersecurity are necessary to identify and mitigate emerging threats, ensuring robust defenses against attacks.

8.6. Countermeasures for Traffic Jams and Noise Pollution

Cybercriminals could exploit AV functionalities to create traffic disruptions and noise pollution. To counter these risks:

- **Decentralized Control Protocols:** Implementing decentralized control systems helps distribute vehicle decision-making, reducing susceptibility to coordinated attacks that could cause traffic jams.
- **Honk Limiting Features and Scheduling:** AVs can incorporate software that restricts honking based on the time of day and location, reducing potential disturbances in residential areas and mitigating risks if systems are compromised.

By implementing these strategies, the AV industry can establish a strong cybersecurity framework that safeguards vehicles against a variety of changing threats, thereby ensuring the safety and security of passengers and urban settings.

9. Statistical Analysis

9.1. The Status of Self-Driving Cars

In 2023, the development of autonomous vehicle (AV) technology is still ongoing, and there are currently only a limited number of self-driving cars being used on U.S. roads. Many of these vehicles are currently undergoing testing to assess and improve their self-driving capabilities. Last year, around 1,400 vehicles, including cars, trucks, and

other types, were undergoing testing by 80 different companies in 36 states throughout the United States. In January 2023, Mercedes-Benz made history by becoming the first automaker in the U.S. to receive government approval for a Level 3 driving feature in Nevada [63]. This achievement marked a significant milestone in the automotive industry. Level 3 automation is of utmost importance as it signifies a remarkable leap forward from Level 2, enabling the vehicle to take care of all driving responsibilities while the driver only needs to step in when required. Despite the relatively small number of autonomous vehicles currently in use, the industry is witnessing significant expansion.

According to market analysis, the global autonomous vehicle (AV) market is expected to experience significant growth in the coming years. The market value is projected to increase from USD 1,921.1 billion in 2023 to USD 13,632.4 billion by 2030, with a compound annual growth rate (CAGR) of 32.3% during this period [64]. In 2022, the Asia-Pacific region emerged as the dominant player in the autonomous vehicle industry, capturing an impressive market share of 50.44%. According to these projections, the autonomous vehicle market is set to experience significant growth, fueled by advancements in technology and the growing acceptance of these vehicles in different regions.

9.2. Consumer Perception and Safety Concerns

Public trust and acceptance are pivotal to the widespread adoption of autonomous vehicles (AVs). While approximately 57% of Americans familiar with self-driving cars are willing to ride in them, and 55% believe that most vehicles will be autonomous by 2029, there remains a significant proportion of the population—43%—who express apprehension about using driverless vehicles [65]. Surveys indicate that safety concerns are at the core of this reluctance. For example, in 2024, AAA reported that about 66% of Americans harbor fears about fully autonomous vehicles due to incidents and perceived risks associated with these systems [66].

Additionally, a World Economic Forum report underscores the role of clear communication and regulatory transparency in enhancing public trust in AV technology. The report emphasizes that as people become more familiar with AV safety measures and technological advancements, their confidence is likely to grow, thereby facilitating a smoother transition to autonomous transportation systems [67]. Understanding and addressing these consumer concerns is crucial for real-world implementation, as public acceptance directly impacts adoption rates and can guide regulatory approaches to foster trust in AV technologies.

9.3. Automated Vehicle Accident Statistics

The safety of self-driving cars is a subject that sparks heated discussions, especially as the technology progresses. In 2022, automakers disclosed around 400 crashes to the National Highway Traffic Safety Administration (NHTSA) involving vehicles equipped with partially automated driver-assist systems. Out of all the incidents, 273 of them involved Tesla vehicles, and interestingly, 70% of these Teslas were using the Autopilot beta feature at the time. It is worth mentioning that 11 of the crashes reported led to severe injuries, while five of them tragically resulted in fatalities. Meanwhile, in 130 reported accidents involving fully autonomous vehicles, there were no injuries in 108 cases, with the majority of incidents being rear-end collisions [65].

9.4. Liability in Autonomous Vehicle Accidents

Assessing responsibility in accidents involving autonomous vehicles introduces an additional level of intricacy to conventional traffic incidents. When a self-driving car is involved in an accident, it becomes essential to determine if the automated driving system was active during the collision and the level of human driver intervention that was anticipated. If there is a suspicion of a flaw in the AV's system, the incident might be subject to product liability law [68]. In such cases, it is necessary to provide evidence of the defect and establish a clear link between the defect and the resulting injury.

9.5. Market and Technological Advancements

The autonomous vehicle industry is making significant strides, with continuous advancements in levels of vehicle automation. Various systems currently exist, with different levels of assistance provided to drivers. At the lowest level, drivers receive only momentary assistance, while at the highest level, the vehicle is capable of handling all driving tasks in any condition. Driver-assist technologies have become a common feature in the majority of new cars. These technologies encompass a range of helpful features such as automatic emergency braking, lane-keeping assistance, and adaptive cruise control [69]. As companies like Mercedes-Benz receive approval to deploy these technologies, more advanced systems, such as Level 3 automation, are becoming increasingly prevalent. These systems allow vehicles to manage most driving tasks independently.

10. Comparative Analysis: Drones vs. Driverless Cars

Technological breakthroughs in autonomous systems have propelled drones and autonomous vehicles to the forefront of contemporary innovation, providing unparalleled advantages in transportation, logistics, and surveillance. Nevertheless, the autonomy that fuels these advantages also brings about substantial cybersecurity vulnerabilities, which give rise to concerns over their possible abuse.

Drones [70] and autonomous vehicles have common technological underpinnings, such as the utilization of GPS for navigation, sensors for environmental perception, and artificial intelligence (AI) for decision-making operations. Both technologies are highly dependent on real-time data processing and external connectivity, rendering them vulnerable to cyber attacks such GPS spoofing, jamming, and unauthorized data interception.

Nevertheless, their operational environments vary greatly. Drones function within airspace and are commonly employed for the purposes of surveillance, delivery, and reconnaissance, rendering them susceptible to abuse in illicit activities such as smuggling, covert surveillance, or even as weaponry [70]. Conversely, autonomous vehicles skillfully maneuver through intricate metropolitan landscapes, prioritizing transit and logistics, where the vulnerabilities to hacks

encompass hijacking, remote control, and the potential for accidents. Understanding these potential threats is crucial for developing appropriate countermeasures.

Given these distinct yet overlapping vulnerabilities, it is essential to explore the specific ways in which both drones and AVs can be exploited by cybercriminals. The following subsections delve into these potential threats, comparing how they manifest in drones and AVs.

10.1. Tracking and Surveillance

Driverless cars can be misused as tools for tracking and surveillance. Since these vehicles are equipped with GPS, cameras, and other sensors, a compromised car can be used to monitor an individual's movements without their knowledge. Cybercriminals or other malicious entities could hijack the car's systems to gather real-time data about the location and activities of the occupants, effectively turning the vehicle into a surveillance tool. While drones are also susceptible to being used for surveillance, their typical use in airspace makes them less likely to be involved in long-term tracking within dense urban environments, where AVs are more commonly found. This kind of misuse raises significant privacy concerns, especially if sensitive personal information is intercepted or recorded.

10.2. Transporting Illegal Items

Driverless cars could be exploited for the unauthorized transportation of illegal goods, such as drugs, weapons, or contraband. Since these vehicles operate autonomously, they could be programmed to follow predetermined routes to specific drop-off locations, reducing the need for human involvement and lowering the risk of detection.

Drones, too, are vulnerable to such misuse, particularly for smuggling goods across borders or into restricted areas. However, the ground-based nature of AVs means they are more likely to be used for covert operations within urban areas. This capability could be particularly appealing to criminal organizations looking to minimize the risks associated with traditional methods of smuggling and transportation.

10.3. Weaponization and Terrorism

In the worst-case scenarios, driverless cars could be weaponized by cybercriminals or terrorists. By hijacking a vehicle's control systems, an attacker could direct the car to drive into a crowded area or a critical piece of infrastructure, such as a bridge or a building. The vehicle could also be loaded with explosives or hazardous materials, turning it into a remote-controlled bomb. While drones have similarly been considered for such malicious purposes due to their aerial capabilities, the ground-based nature of autonomous vehicles allows them to carry larger payloads and target densely populated urban environments more effectively. The potential for such attacks highlights the importance of securing the software and communication systems of autonomous vehicles to prevent their use in acts of terrorism.

10.4. Coordinated Cyberattacks

Driverless cars, like drones, are part of a broader network of interconnected systems, including other vehicles, traffic management systems, and cloud-based services. This interconnectivity increases the risk of coordinated cyberattacks [71], where multiple vehicles are compromised and used in concert to achieve a malicious goal. Such attacks could disrupt urban infrastructure, create widespread panic, or even be used as a tool for cyber warfare. The ability to remotely control large numbers of autonomous vehicles presents a unique challenge to cybersecurity professionals, who must develop strategies to detect and neutralize such threats before they can cause harm. For instance, the Ukrainian government has developed advanced drone systems capable of coordinated attacks [72]. These drones can intercommunicate, autonomously decide on targets, and collect intelligence at speeds surpassing human capabilities. This development illustrates the potential for autonomous systems to be used in coordinated actions, highlighting the need for robust cybersecurity measures in both military and civilian autonomous vehicle networks.

11. Societal Impacts

11.1. Transportation Accessibility

Driverless cars have the potential to significantly improve transportation accessibility for individuals with disabilities, the elderly, and those who cannot drive. Autonomous vehicles (AVs) can provide on-demand mobility services, enhancing independence and quality of life for these populations. For example, AVs could offer new mobility options to the approximately 49 million Americans over the age of 65 and the 53 million people with some type of disability. This increase in accessibility could result in a 14% rise in vehicle miles traveled, potentially worsening congestion but significantly improving individual mobility and independence [73]. Furthermore, automated vehicles could create new employment opportunities for about 2 million people with disabilities in the United States [74].

11.2. Urban Planning and Traffic Management

Autonomous vehicles are expected to have a profound impact on urban planning and traffic management. By reducing the need for extensive parking facilities and enabling more efficient use of road space, AVs can contribute to better urban environments. Improved traffic management through vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication can reduce congestion and optimize traffic flow. Platooning, where vehicles travel closely together, can increase road capacity and reduce traffic delays [73]. Additionally, shared mobility services facilitated by AVs can support car-sharing concepts, reducing the number of privately owned vehicles and promoting more sustainable urban transport models [74].

11.3. Environmental Considerations

The adoption of electric autonomous vehicles can significantly contribute to environmental sustainability by reducing greenhouse gas emissions and improving air quality. Efficient driving patterns and reduced congestion lead to lower fuel consumption and environmental impact. According to the U.S. Chamber of Commerce, autonomous vehicles

have the potential to prevent 1.4 million accidents and 12,000 fatalities annually, resulting in substantial economic savings of $94 billion [75]. Moreover, the shift towards shared autonomous vehicles can further reduce the overall number of vehicles on the road, decreasing the carbon footprint associated with manufacturing and operating personal vehicles.

11.4. Job Displacement and Economic Implications

The transition to driverless cars may lead to job displacement in industries such as trucking, taxi services, and public transportation. In the United States, approximately 2.9% of workers are employed in driving occupations, with more than 4 million potentially affected by the adoption of AVs. This could result in an estimated annual income loss of around $180 billion for these workers [73]. However, the rise of autonomous vehicles also presents opportunities for economic growth, including new jobs in technology development, infrastructure upgrades, and mobility services. The motor insurance industry, worth over $300 billion annually in the U.S., may also experience significant changes as AVs reduce the frequency and severity of accidents, potentially lowering insurance premiums for consumers [73].

12. Regulatory Landscape

12.1. Government Initiatives and Policies

Governments worldwide are actively developing initiatives and policies to support the deployment of autonomous vehicles (AVs). These efforts include funding for research and development, establishing testing and certification standards, and creating frameworks for AV integration into existing transportation systems. For instance, in the United States, the National Highway Traffic Safety Administration (NHTSA) continues to lead federal efforts, focusing on creating guidelines that ensure safety while promoting innovation. Various states are also pioneering regulatory efforts by passing laws and executive orders to facilitate AV testing and deployment, highlighting the importance of harmonized federal regulations to avoid a fragmented regulatory landscape.

12.2. Legal Framework for Autonomous Vehicles

Establishing a robust legal framework is essential for the safe and effective deployment of AVs. Key components of this framework include:

1. **Defining Liability**: Clarifying liability in the event of accidents involving AVs is crucial. This involves determining the responsibility between manufacturers, software developers, and vehicle operators. Some states, like California, Nevada, and Florida, have implemented requirements for manufacturers to hold significant insurance policies or bonds to cover potential damages.

2. **Setting Standards for Performance and Safety**: Governments are establishing standards for vehicle performance and safety to ensure that AVs can operate safely on public roads. These standards cover aspects such as sensor accuracy, fail-safe mechanisms, and emergency response capabilities.

3. **Ensuring Data Privacy Compliance**: As AVs collect and process vast amounts of data, ensuring compliance with data privacy regulations is essential to protect user information and maintain public trust.

12.3. International Standards and Collaboration

International collaboration and the establishment of global standards are crucial for the interoperability and scalability of autonomous vehicle technology. Organizations such as the International Organization for Standardization (ISO) and SAE International play key roles in developing these standards. For example, the European Union has finalized a legal framework for fully automated vehicles, creating binding regulations that include comprehensive safety and performance criteria. Collaborative efforts are also evident in cross-border initiatives, where countries work together to create harmonized regulatory approaches, enabling seamless operation of AVs across different jurisdictions.

12.4. Comparative Regulatory Frameworks in Europe and Asia

European and Asian countries have developed distinctive regulatory frameworks that address the deployment and operation of AVs, reflecting diverse regional priorities and challenges. In the **European Union (EU)**, regulations such as the General Safety Regulation mandate that all new vehicles include advanced driver assistance systems (ADAS) by 2022, encompassing features like lane-keeping and automated emergency braking. Furthermore, *UN Regulation 157* on Automated Lane Keeping Systems (ALKS) permits Level 3 AVs on public roads, provided they meet strict safety, cybersecurity, and software update requirements. Germany, a leader within the EU, enacted a law in 2021 allowing Level 4 AVs to operate in specified public areas, setting a precedent for comprehensive AV legislation [76].

In the **United Kingdom**, efforts to position itself as an AV innovation hub are underscored by significant investments exceeding £200 million in AV testing and research infrastructure. The UK's *Code of Practice for Testing Automated Vehicles* provides guidelines on liability, insurance, and data protection, facilitating AV trials on public roads. By 2025, the UK aims to establish a legal framework to support widespread AV deployment, focusing on safety standards and clarifying manufacturer and user responsibilities [77].

In **Asia**, Japan has prioritized AV integration into public transport, especially in rural areas, and revised the *Road Transport Vehicle Law* to permit Level 3 AVs on public roads. Japan is actively preparing for the *2025 World Expo* in Osaka, where Level 4 AVs are expected to be deployed [78]. China, another major player, has developed city-specific policies, allowing for Level 4 testing in cities like Beijing and Shanghai. The *Beijing Autonomous Driving Policy* includes data security requirements, emphasizing compliance with national laws like the *Cybersecurity Law* and the *Personal Information Protection Law (PIPL)* [79]. South Korea, aspiring to be an AV industry leader, has implemented the *Framework Act on Intelligent Robots* to regulate AVs, along with establishing *K-City*, a large-scale AV testing facility simulating urban and highway environments [80].

13. Results

In this section, the primary findings of the study are presented, with a particular emphasis on the comparative analysis of cybersecurity threats facing autonomous vehicles (AVs) and drones. By examining the common vulnerabilities between these two technologies, this research sheds light on potential threats that AVs may encounter, drawing parallels from the established cybersecurity issues in drone systems. This cross-technology perspective provides valuable insights into emerging risks and underscores the importance of proactive threat mitigation.

Additionally, a comparison with recent related works is included to underscore the unique contributions of this study. Table 2 summarizes how this research diverges from existing literature by specifically focusing on emerging and future cybersecurity threats in AVs, inferred from drone vulnerabilities. This approach offers a forward-looking view on cybersecurity challenges, providing a foundation for understanding and anticipating potential misuse of AV technology.

Table 2: Threats in Autonomous Vehicles with Status

Threat	Potential Impact on AVs	Status
Remote Exploits	Unauthorized control over AV functions, posing safety risks	Existing
GPS Spoofing	Misleading AV navigation, causing route deviations	Emerging
Adversarial ML Attacks	Misinterpretation of road signs or obstacles, leading to unsafe decisions	Emerging
Supply Chain Attacks	Risk of tampered AV components before deployment	Emerging
Ransomware	Potential to incapacitate AV systems until payment is made	Possible Future
Denial of Service (DoS)	System overload, disrupting AV operations and real-time data processing	Possible Future
Noise Pollution	Excessive AV honking and increased vehicle numbers could heighten urban noise levels	Possible Future
Traffic Congestion	More AVs on roads could worsen traffic if not managed effectively	Possible Future
Risks in AV Charging Infrastructure	Cyberattacks on charging stations, grid strain, and bottlenecks at charging sites	Possible Future
Tracking and Surveillance	Unauthorized tracking using AVs for long-term monitoring within urban areas	Emerging
Transporting Illegal Items	Use of AVs for transporting contraband autonomously within urban settings	Possible Future
Weaponization and Terrorism	Hijacking AVs for attacks or as remote-controlled bombs targeting urban areas	Possible Future
Coordinated Cyberattack	Large-scale attacks involving multiple AVs to disrupt traffic or emergency response systems	Possible Future

14. Limitations

While this paper offers a broad survey of autonomous vehicle (AV) technology, it is limited by its high-level perspective, which does not delve into specific implementation details. The generalized nature of this study may not fully reflect the unique technological approaches of different manufacturers, thereby limiting the applicability of its findings across various AV systems. Additionally, due to the rapid pace of advancements in AV technology, some insights presented here may quickly become outdated as new innovations in sensors, AI algorithms, and cybersecurity measures continue to evolve.

Furthermore, the study focuses mainly on technological and cybersecurity challenges, with less emphasis on the ethical and social implications of AV adoption. Complex issues, such as the potential impact on employment, ethical considerations in AI-driven decision-making, and regional differences in regulatory readiness, are acknowledged but not explored in depth. Future research could address these areas by conducting case studies, longitudinal analyses, and interdisciplinary investigations that incorporate ethical, social, and regional dimensions, providing a more holistic view of the challenges and opportunities surrounding autonomous vehicles.

15. Conclusion

Autonomous vehicles (AVs) signify a profound paradigm shift within the transportation sector, bearing both theoretical and practical ramifications. From a theoretical standpoint, AVs provide insights into the interplay of artificial intelligence, sensor integration, and real-time decision-making within complex systems. On a practical level, they present promising strategies to enhance road safety, alleviate traffic congestion, and mitigate the environmental consequences associated with conventional transportation. This study illustrates how the convergence of advanced sensors, AI algorithms, and vehicle-to-everything (V2X) communication technologies is advancing the AV landscape.

A key contribution of this research is the delineation of essential technologies and the examination of cybersecurity challenges that AVs encounter. By analyzing these elements, this study highlights the imperative need for robust cybersecurity frameworks designed to safeguard AV systems against potential threats. Furthermore, the study underscores the necessity for comprehensive regulatory frameworks that can foster public trust and support the widespread adoption of AV technology.

From an applied perspective, AVs are poised to deliver numerous benefits. They hold the potential to drastically reduce road fatalities attributable to human error, which remains a leading cause of vehicular accidents worldwide. In addition, AVs can enhance traffic management by optimizing routing and facilitating communication between vehicles and infrastructure, thereby promoting more efficient road usage. Moreover, the shift towards electric AVs offers considerable potential for reducing greenhouse gas emissions and supporting broader environmental sustainability goals.

Nevertheless, this research acknowledges certain limitations. Although it provides a comprehensive overview of AV technology and its associated challenges, the study does not delve into the specific implementation variations across different AV manufacturers. Additionally, given the rapid pace of technological advancements in this domain, the findings presented are susceptible to obsolescence. Future research could address these limitations by conducting longitudinal studies to track the ongoing evolution of AV technologies.

Looking forward, several avenues for future research merit attention. First, there is a pressing need for the development of standardized cybersecurity protocols that address the specific challenges posed by real-time data exchange and autonomous decision-making in AVs. Second, further research should explore the socio-economic impacts of AV proliferation, particularly concerning potential job displacement within the driving sector. Finally, interdisciplinary studies that integrate perspectives from urban planning, ethics, and AI safety are essential to comprehensively under-

stand how AVs will reshape transportation infrastructure and societal frameworks.

In conclusion, autonomous vehicles offer a transformative opportunity to develop transportation systems that are safer, more efficient, and more sustainable. Addressing the current challenges will necessitate ongoing collaboration among policymakers, industry stakeholders, and the research community. By surmounting these obstacles, AVs have the potential to become a fully integrated aspect of modern society, fundamentally altering the way we navigate our world for generations to come.

References

[1] T. Kanade, C. Thorpe, W. Whittaker, "Autonomous land vehicle project at cmu", "Proceedings of the 1986 ACM Fourteenth Annual Conference on Computer Science", CSC '86, pp. 71–80, ACM, New York, NY, USA, 1986, doi:10.1145/319838.319850.

[2] Mobileeye, "History of autonomous vehicles: From renaissance to reality", https://shorturl.at/asq4I, 2023, accessed: 2024-07-05.

[3] Arrow, "History of self-driving cars", https://www.arrow.com/en/research-and-events/articles/the-history-of-self-driving-cars, accessed: 2024-07-05.

[4] S. Magazine, "The national automated highway system that almost was", https://shorturl.at/UbL7x, 2013, accessed: 2024-07-05.

[5] Google, "Waymo: On the road", https://waymo.com/ontheroad/, 2017, accessed: 2024-07-05.

[6] G. V. Research, "Autonomous vehicle market to reach $214.32bn by 2030", https://www.grandviewresearch.com/industry-analysis/autonomous-vehicle-market, 2023, accessed: 2024-07-05.

[7] McKinsey and Company, "Autonomous vehicles moving forward: Perspectives from industry leaders", 2024, article.

[8] K. Bian, G. Zhang, L. Song, "Security in use cases of vehicle-to-everything communications", "2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)", pp. 1–5, 2017, doi:10.1109/VTCFall.2017.8288208.

[9] F. B. Insights, "Autonomous vehicle market size, share, trends | report [2030]", https://www.fortunebusinessinsights.com/autonomous-vehicle-market-102020, 2023, accessed: 2024-07-05.

[10] P. Research, "Autonomous vehicle market size to hit usd 2,752.80 bn by 2033", https://www.precedenceresearch.com/autonomous-vehicle-market, 2023, accessed: 2024-07-05.

[11] Precedence Research, "Autonomous vehicle market size to hit usd 2,752.80 bn by 2033", https://www.precedenceresearch.com/autonomous-vehicle-market#:~:text=The%20global%20autonomous%20vehicle%20market,USD%2059.92%20billion%20in%202023, 2024, last updated: June 2024.

[12] W. H. Organization, "Global status report on road safety 2018", *World Health Organization*, 2018, accessed: 2024-07-05.

[13] WHO, "Global status report on road safety 2023", *World Health Organization*, 2023, accessed: 2024-07-05.

[14] W. H. Organization, "Road traffic injuries", *World Health Organization*, 2023, accessed: 2024-07-05.

[15] Spectrum, "Autonomous vehicles can make all cars more efficient", https://shorturl.at/9ht7y, 2018, accessed: 2024-07-05.

[16] McKinsey, Company, "The future of autonomous vehicles (av)", https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-future-of-autonomous-vehicles-av, 2023, accessed: 2024-07-05.

[17] ScienceDirect, "The impacts of connected autonomous vehicles on mixed traffic flow: A review", https://www.sciencedirect.com/science/article/pii/S2352146520305170, 2020, accessed: 2024-07-05.

[18] N. Geographic, "Energy implications of autonomous vehicles: Imagining the possibilities", https://www.nationalgeographic.com/environment/article/energy-implications-of-autonomous-vehicles, 2019, accessed: 2024-07-05.

[19] B. C. Group, "Can self-driving cars stop the urban mobility meltdown?", https://shorturl.at/F0d5d, 2020, accessed: 2024-07-05.

[20] R. Du, *et al.*, "Effective urban traffic monitoring by vehicular sensor networks", *IEEE Transactions on Vehicular Technology*, vol. 64, no. 1, pp. 273–286, 2014, doi:10.1109/TVT.2014.2349320.

[21] Y. Li, *et al.*, "Vehicle detection based on the and–or graph for congested traffic conditions", *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 984–993, 2018, doi:10.1109/TITS.2018.2871040.

[22] R. F. Foundation, "Self-driving cars: The impact on people with disabilities", *Ruderman White Paper*, 2017.

[23] R. Corporation, "Self-driving vehicles offer potential benefits, policy challenges for lawmakers", *RAND Corp*, 2014.

[24] N. C. Onat, J. Mandouri, M. Kucukvar, *et al.*, "Rebound effects undermine carbon footprint reduction potential of autonomous electric vehicles", *Nature Communications*, vol. 14, p. 6258, 2023, doi:10.1038/s41467-023-41992-2, received 01 May 2023, Accepted 22 September 2023, Published 06 October 2023.

[25] A. J. G. Rodríguez, N. J. Barón, J. M. G. Martínez, "Validity of dynamic capabilities in the operation based on new sustainability narratives on nature tourism smes and clusters", *Sustainability*, vol. 12, no. 3, p. 1004, 2020, doi:10.3390/su12031004.

[26] Y. Zhang, Z. Wang, H. Wang, F. Blaabjerg, "Artificial intelligence-aided thermal model considering cross-coupling effects", *IEEE Transactions on Power Electronics*, vol. 35, no. 10, pp. 9998–10002, 2020, doi:10.1109/TPEL.2020.2980240.

[27] KPMG, "Connected and autonomous vehicles - readiness index", https://t.ly/_tAW_, 2020, accessed: 2024-07-05.

[28] SMMT, "Connected and autonomous vehicles: The global race to market", https://tinyurl.com/4kdakunk, 2021, accessed: 2024-07-05.

[29] R. Domínguez, E. Onieva, J. Alonso, J. Villagra, C. González, "Lidar based perception solution for autonomous vehicles", "Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on", pp. 790–795, 2011.

[30] E. Hasch, R. Topak, R. Schnabel, T. Zwick, R. Weigel, C. Waldschmidt, "Millimeter-wave technology for automotive radar sensors in the 77 ghz frequency band", *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 3, pp. 845–860, 2012, doi:10.1109/TMTT.2011.2178427.

[31] M.-H. Lee, Y.-J. Chen, T. Li, "Sensor fusion design for navigation and control of an autonomous vehicle", "2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)", pp. 2209–2214, 2011, doi:10.1109/ICSMC.2011.6084000.

[32] Q. Li, L. Chen, M. Li, S.-L. Shaw, A. Nuchter, "A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios", *IEEE Transactions on Vehicular Technology*, vol. 63, no. 2, pp. 540–555, 2014, doi:10.1109/TVT.2013.2289913.

[33] B. Padmaja, C. Moorthy, N. Venkateswarulu, *et al.*, "Exploration of issues, challenges and latest developments in autonomous cars", *Journal of Big Data*, vol. 10, no. 61, 2023, doi:10.1186/s40537-023-00701-y, published: 06 May 2023.

[34] L. Hobert, A. Festag, I. Llatser, L. Altomare, F. Visintainer, A. Kovacs, "Enhancements of v2x communication in support of cooperative autonomous driving", *IEEE Communications Magazine*, vol. 53, pp. 64–70, 2015, doi:10.1109/MCOM.2015.7105641.

[35] S. A. Centers, "The six levels of autonomous driving", https://www.schaeferautobody.com/the-six-levels-of-autonomous-driving/#:~:text=Conditional%20Automation%20(Level%203)%20Vehicles%20at%20this,to%20monitor%20for%20changes%20in%20those%20conditions, 2019, accessed: October 11, 2024.

[36] A. Greenberg, "Hackers remotely kill a jeep on the highway—with me in it", https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/, 2015, accessed: 2024-08-26.

[37] The Guardian, "Uber's joe sullivan faces trial over data breach cover-up", *The Guardian*, 2022, accessed: 2024-08-30.

[38] H. Shin, D. Won, S. Kim, "Gps spoofing attack on autonomous vehicles and its countermeasure", *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1738–1748, 2019, doi:10.1109/TITS.2019.2896358.

[39] J. Kenney, "Jamming attack vulnerabilities in v2x communication systems", "IEEE Vehicular Technology Conference (VTC2021-Fall)", 2021, doi:10.1109/VTCFall2021.2019.9198881.

[40] A. Hamid, P. Lin, R. Hussain, "Relay attacks in vehicular networks: Analysis, impact, and solutions", *IEEE Communications Magazine*, vol. 60, no. 4, pp. 79–85, 2022, doi:10.1109/MCOM.2022.3056782.

[41] Q. Yan, D. Zhang, "Sensor spoofing attacks on autonomous vehicles: A review of the vulnerabilities and mitigation strategies", *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2835–2856, 2020, doi:10.1109/COMST.2020.2976336.

[42] I. Foster, K. Koscher, "Tesla model s: A case study in software security and over-the-air updates", *Journal of Cybersecurity*, vol. 6, no. 1, pp. 1–15, 2018, doi:10.1093/cybsec/tyy010.

[43] J. Petit, S. Shladover, "Mitm attacks on connected vehicles: Techniques, impact, and countermeasures", *IEEE Vehicular Technology Magazine*, vol. 15, no. 2, pp. 70–77, 2020, doi:10.1109/MVT.2020.2983621.

[44] M. Groll, S. R. Weller, "Denial-of-service attacks in autonomous vehicle networks: A comprehensive analysis", *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 3, pp. 400–412, 2019, doi:10.1109/TIV.2019.2937765.

[45] T. B. Brown, D. Mane, A. Roy, M. Abadi, J. Gilmer, "Adversarial patch", https://arxiv.org/abs/1907.07736, 2019, accessed: 2024-08-26.

[46] U.S. News and World Report, "Car dealerships are being disrupted by a multi-day outage after cyberattacks on software supplier", https://shorturl.at/1bgnQ, 2024, accessed: 2024-08-30.

[47] S. C. Nayak, V. Tiwari, B. K. Samanthula, "Review of ransomware attacks and a data recovery framework using autopsy digital forensics platform", "2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)", pp. 0605–0611, 2023, doi:10.1109/CCWC57344.2023.10099169.

[48] L. Eliot, "Here's how ransomware is going to fiendishly impede ai self-driving cars", https://shorturl.at/ptDMH, 2021, accessed: 2024-08-30.

[49] Nahla Davies, "Keyless car theft", https://shorturl.at/d5PLL, 2023, accessed: 2024-08-30.

[50] Alan J, "Bluetooth relay attacks", https://thecyberexpress.com/tesla-ultra-wideband-vulnerable-relay-attacks/, 2024, accessed: 2024-08-30.

[51] David Strom, "Ev charging stations still riddled with cybersecurity vulnerabilities", https://shorturl.at/vpv6y, 2024, accessed: 2024-08-30.

[52] Chris Isidore, "Driverless cars flood san francisco: What happens when things go wrong?", https://www.cnn.com/2023/08/14/business/driverless-cars-san-francisco-cruise/index.html, 2023, accessed: 2024-08-30.

[53] J. Marcus, "San francisco residents fed up with self-driving car honking", https://shorturl.at/zQT5p, 2024, accessed: 2024-10-10.

[54] I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial examples", *arXiv preprint arXiv:1412.6572*, 2015, adversarial training improves model robustness by training AI models on adversarial examples.

[55] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks", *IEEE Symposium on Security and Privacy*, 2016, defensive distillation smooths decision boundaries to mitigate the impact of adversarial inputs.

[56] W. Xu, D. Evans, Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks", *arXiv preprint arXiv:1704.01155*, 2018, feature squeezing reduces input complexity to counter adversarial attacks.

[57] X. Huang, M. Kwiatkowska, S. Wang, M. Wu, "Safety verification of deep neural networks", *arXiv preprint arXiv:1610.06940*, 2017, robustness verification through formal methods ensures AI model resilience against adversarial perturbations.

[58] IEEE Smart Cities, "Secure boot and hardware roots of trust in autonomous vehicles", https://smartcities.ieee.org/security/autonomous-vehicle-security, 2023, accessed: 2023-10-11.

[59] IBM - United States, "Ibm blockchain technology", https://www.ibm.com/blockchain/supply-chain, 2023, accessed: 2023-10-11.

[60] Y. Rahayu, T. A. Rahman, R. Ngah, P. Hall, "Ultra wideband technology and its applications", "2008 5th IFIP International Conference on Wireless and Optical Communications Networks (WOCN '08)", pp. 1–5, 2008, doi:10.1109/WOCN.2008.4542537.

[61] S. Ibrokhimov, K. L. Hui, A. Abdulhakim Al-Absi, h. j. lee, M. Sain, "Multi-factor authentication in cyber physical system: A state of art survey", "2019 21st International Conference on Advanced Communication Technology (ICACT)", pp. 279–284, 2019, doi:10.23919/ICACT.2019.8701960.

[62] D. Zelle, C. Krauß, H. Strauß, K. Schmidt, "On using tls to secure in-vehicle networks", "Proceedings of the 12th International Conference on Availability, Reliability and Security", ARES '17, Association for Computing Machinery, New York, NY, USA, 2017, doi:10.1145/3098954.3105824.

[63] Mercedes-Benz Group, "Drive pilot: Nevada becomes first us state to approve sae level 3 system for mercedes-benz", https://group.mercedes-benz.com/innovation/product-innovation/autonomous-driving/drive-pilot-nevada.html, 2024, accessed: 2024-08-30.

[64] Fortune Business Insights, "Autonomous vehicle market size, share and covid-19 impact analysis, by component (hardware and software), by type (passenger car and commercial vehicle), and regional forecast, 2023-2030", https://www.fortunebusinessinsights.com/autonomous-vehicle-market-109045, 2024, accessed: 2024-08-30.

[65] KNR Legal, "Self-driving car accident statistics", https://www.knrlegal.com/car-accident-lawyer/self-driving-car-accident-statistics/, 2024, accessed: 2024-08-30.

[66] AAA, "2024 aaa survey on autonomous vehicles", https://newsroom.aaa.com/2024/05/aaa-survey-autonomous-vehicles/, 2024, accessed: 2024-10-10.

[67] W. E. Forum, "Which trends are driving the autonomous vehicles industry?", https://www.weforum.org, accessed: 2024-07-05.

[68] Byrd Davis Alden and Henrichson, LLP, "Who is liable when a self-driving car causes a crash?", https://shorturl.at/yZ7nY, 2024, accessed: 2024-08-30.

[69] Alliance for Automotive Innovation, "Autonomous vehicles", https://www.autosinnovate.org/initiatives/innovation/autonomous-vehicles, 2024, accessed: 2024-08-30.

[70] S. C. Nayak, B. K. Samanthula, V. Tiwari, "Investigating drone data recovery beyond the obvious using digital forensics", "2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)", pp. 0254–0260, 2023, doi:10.1109/UEMCON59035.2023.10315995.

[71] Cybersecurity and Infrastructure Security Agency (CISA), "Action guide for cybersecurity: Critical control actions", https://www.cisa.gov/sites/default/files/2022-11/Action%20Guide%20CCA%20508%20FINAL%2020190905.pdf, 2022, accessed: 2024-08-30.

[72] Kyiv Post, "Ukraine's tech hub develops ai-driven drone swarms to combat russian forces", https://www.kyivpost.com/post/34777, 2024, accessed:2024-08-30.

[73] F. Klaver, "The economic and social impacts of fully autonomous vehicles", https://www.compact.nl, 2021, accessed: 2024-07-05.

[74] Bosch, "Impact of self-driving cars on society", https://www.bosch.com, accessed: 2024-07-05.

[75] U. Chamber, "New u.s. chamber report on economic and social benefits of autonomous vehicles", https://www.uschamber.com, accessed: 2024-07-05.

[76] European Commission, "Commission adopts new rules to support safe deployment of automated and connected vehicles in the eu", https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4312, 2022, accessed: 2023-10-11.

[77] United Nations Economic Commission for Europe (UNECE), "Un regulation on automated lane keeping systems (alks) extended to trucks, buses and coaches", https://unece.org/automated-lane-keeping-system-alks, 2021, https://unece.org/automated-lane-keeping-system-alks.

[78] Japan Times, "Kishida administration to boost self-driving car development in japan", https://www.japantimes.co.jp/news/2024/08/01/japan/kishida-self-driving-cars-boost/, 2024, accessed: 2023-10-11.

[79] T. Staff Reporter, W. Ke, "China accelerates autonomous driving via multiple pilot cities, with beijing, shanghai, wuhan leading the charge", https://t.ly/sMNut, 2024, accessed: 2023-10-11.

[80] Applied Intuition, "How the korean police science institute uses simian and basis to validate av stack performance", https://www.appliedintuition.com/news/korea-police-science-institute, 2023, accessed: 2023-10-11.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Vaibhavi Tiwari holds a MicroMasters credential in Data and Statistics from the Massachusetts Institute of Technology (MIT). She completed her bachelor's degree in Computer Applications from an esteemed institution in India and earned a double master's degree: one in Computer Applications from NITK Surathkal and another in Computer Science from Montclair State University. With over eight years of experience in healthcare technology, Vaibhavi specializes in data security, AI integration, and Cloud based technologies. Currently, she is the Head of Technology and Engineering at MyGreen Health, where she leads the development of innovative healthcare solutions.

Vaibhavi's research lies at the intersection of cybersecurity, healthcare, and big data. She has published extensively in IEEE conferences, with recent works including "Review of Ransomware Attacks and a Data Recovery Framework using Autopsy Digital Forensics Platform" and "Investigating Drone Data Recovery Beyond the Obvious Using Digital Forensics." These studies underscore her commitment to tackling security challenges within healthcare and other critical sectors. Her contributions emphasize proactive threat identification and mitigation strategies, showcasing her expertise in securing complex, data-driven environments. Her accolades include being a certified Globee Awards judge, which highlights her contributions to technology and business evaluation. Vaibhavi's work spans research, professional development, and community service, such as her mentorship in the Women in Big Data program and her active volunteer efforts for the ISEF.

Received: 30 August 2024, Revised: 14 October, 2024, Accepted: 15 October, 2024, Online: 24 October, 2024

DOI: https://doi.org/10.55708/js0310005

Evaluation of Equivalent Aacceleration Factors of Repairable Systems in a Fleet: a Process-Average-Based Approach

Renyan Jiang [1, 2], Kunpeng Zhang [1, 2], Xia Xu [2, 3], Yu Cao [1, 2*]

[1] China International Science & Technology Cooperation Base for Laser Processing Robotics, Wenzhou University, Wenzhou 325035, China.

[2] Zhejiang Provincial Innovation Center of Laser Intelligent Equipment Technology, Wenzhou, 325000, China.

[3] Penta laser (Zhejiang) Co., Ltd., Wenzhou 325000, China.

Corresponding Author: Yu Cao, Wenzhou University, E-mail: yucao@wzu.edu.cn

ABSTRACT: Research on repairable systems in a fleet is mainly concerned with modelling of the failure times using point processes. One important issue is to quantitatively evaluate the heterogeneity among systems, which is usually analyzed using frailty models. Recently, a fleet heterogeneity evaluation method is proposed in the literature. This method describes the heterogeneity with the relative dispersion of equivalent acceleration factors (EAFs) of systems, which is defined as the ratio of the mean times between failures (MTBFs) of a system and a reference system. A main drawback of this method is that the MTBFs of a specific system and the reference system are estimated at different times while the MTBF estimated at different time can be different. This paper aims to address this issue by proposing an improved method. The proposed method uses an "average process" as the reference process and estimates the MTBFs of systems and the reference system at a common time point. This leads to more robust MTBF estimates. Three datasets are analyzed to illustrate the proposed method and its superiority.

KEYWORDS: Repairable system; reliability modeling; MTBF; equivalent acceleration factor; fleet heterogeneity

1. Introduction

The literature on repairable systems is vast and mainly concerned with modelling of the failure times using point processes [1-2]. The repairable systems that most of the literature deals with can be roughly divided into two categories: (a) multi-component repairable systems composed by non-identical components, and (b) repairable systems in a fleet composed by identical or similar units. In this paper, we focus on the repairable systems in a fleet.

The reliability modeling of repairable systems in a fleet involves a number of issues, including:

- Failure trend analysis [3-4] and failure pattern analysis [5-6]. The former deals with the behavior of the failure occurrence rate or failure intensity, e.g., increasing, decreasing or non-monotonic over time; and the latter aims to identify typical shape types of failure intensity function.
- Fleet heterogeneity evaluation and/or modeling [7-10]. Basic models are frailty models [11-13] and the equivalent acceleration factor model can be viewed as an exception [14].
- Reliability model selection [15-16] and development of models and/or methods for modeling the data with different types of information or covariates (e.g., failure causes and modes) and different types of censoring [17-26].
- Maintenance quality evaluation [27], maintenance decision optimization [14, 23, 28], and prediction of number of failures [29-30].

The focus of this paper is on the evaluation of heterogeneity of repairable systems in a fleet.

Different from the frailty models, Jiang et al. [14] recently propose a heterogeneity evaluation method for repairable systems in a fleet. It estimates the mean time between failures (MTBFs) of individual systems at the right end point of the observation window of each system

and defines the equivalent acceleration factor (EAF) of a specific system as the ratio between the MTBFs of this system and a reference system, which is the one whose failure number is the largest among all the systems. The fleet heterogeneity is measured by the ratio of the sample average and standard deviation of EAFs of all the systems. The smaller this ratio is, the greater the heterogeneity of the fleet is. Generally, the observation windows of systems are different so that the MTBFs of different systems are estimated at different time points while the MTBF estimated at different time is different and can be considerably overestimated when the failure number is small. As a result, the EAFs estimated from the above-mentioned heterogeneity evaluation method can be unreliable. This paper aims to address this research gap through proposing an improved method to estimate the EAFs and MTBFs of systems. The proposed method uses an "average system" as the reference system or process, whose mean cumulative function (MCF) equals to the MCF of the fleet. The EAF of a system is estimated as the ratio between the MCFs of the system and fleet (or reference system) at the right end point of the observation window of the system; and the MTBF of the system is estimated as the product of the EAF and fleet MTBF. A unique advantage of using the average process as the reference process is that any data point on the MCF of a specific unit always corresponds to a data point on the MCF of the average process. The reference process defined in [14] does not have such a property. This property makes it possible to assess the MTBFs at the same time. Three datasets are analyzed to illustrate the proposed method and its superiority.

The paper is organized as follows. The proposed method is presented in Section 2, compared with the frailty model in Section 3, and illustrated in Section 4. The paper is concluded in Section 5.

2. Proposed method

2.1. Nelson MCF estimator

Consider a fleet of nominally identical repairable systems, and each system is called a unit. Failure point processes of units are randomly censored on the right. Let t_{ij} denote the time to the jth failure of the ith unit, and $x_{ij} = t_{ij} - t_{i,j-1}$ $(t_{i,0} = 0)$ denote the inter-failure time. The failure point process data are given by

$$\{t_{ij} \le \tau_i, 1 \le i \le n, 1 \le j \le n_i\} \tag{1}$$

where τ_i is the censored time of the ith unit, and n_i is the failure number of the ith unit. The fleet is single censoring if $\tau_1 = \tau_2 = \cdots = \tau_n$; otherwise, it is multiple censoring. The total failure number of the fleet is $N = \sum_{i=1}^{n} n_i$, and the censoring time of the fleet is defined as

$$\tau = \max_i(\tau_i). \tag{2}$$

The total time on test (TTT, or T_3 for short) of the fleet at τ is $T_3(\tau) = \sum_{i=1}^{n} \tau_i$, and the fleet MTBF evaluated at τ is estimated as

$$\mu_F = T_3(\tau)/N. \tag{3}$$

A failure point process can be characterized by the MCF of the fleet, which is the mean cumulative number of failures per unit. The MCF can be used to predict the total number or cost of repairs of the fleet in a future period and to determine an optimum retirement age of a system [7]. To estimate the MCF of the fleet, the data given by (1) are sorted in an ascending order. Let t_k $(1 \le k \le N)$ denote the k th smallest value of t_{ij} 's, and $s(t_k)$ denote the number of units at risk at time t_k. For a single censoring fleet, $s(t_k) = n$; for a multiple censoring fleet, $s(t_k)$ decreases with k and tends to 1. The Nelson MCF estimator is given by [7-8]

$$M_N(t_k) = M_N(t_{k-1}) + \frac{1}{s(t_k)},\ t_0 = 0, M_N(0) = 0. \tag{4}$$

Equation (4) is a staircase function with $M_N(t_k^-) = M_N(t_{k-1}^+)$. That is, the MCF at t_k has two values. To make the MCF at t_k unique, a smoothed MCF is defined as

$$M(t_k) = \frac{M_N(t_{k-1}) + M_N(t_k)}{2}. \tag{5}$$

An MCF-based fleet MTBF can be estimated as

$$\mu_{FN} = \frac{\tau}{M(\tau)}. \tag{6}$$

Thus, we have two the MTBF estimates obtained from (3) and (6), respectively. For a single censoring fleet, we have $T_3(\tau) = n\tau$ and $M(\tau) = N/n$. From (6) yields

$$\mu_{FN} = \frac{\frac{T_3(\tau)}{n}}{\frac{N}{n}} = \frac{T_3(\tau)}{N} = \mu_F. \tag{7}$$

In this case, (6) is consistent with (3). For a multiple censoring fleet, the MTBF estimated from (6) is generally different from the MTBF estimated from (3). As mentioned by Nelson [7], the Nelson MCF estimator for large t is not robust for a multiple censoring fleet since $s(t)$ decreases as t increases. Therefore, the fleet MTBF estimated by (3) is preferred.

2.2. Robust MCF estimator

To improve the Nelson MCF estimator, a robust estimator is recently proposed by [14]. Specific details are outlined as follows.

Suppose that an ordered dataset $(t_k, 1 \le k \le N)$ is collected from a multiple censoring fleet with n units. An equivalent operating time t_k^* associated with t_k is defined by

$$t_k^* = \frac{T_3(t_k)}{n}. \tag{8}$$

Define an equivalent single censoring fleet with n units, whose censoring time is defined as

$$\tau^* = \frac{T_3(\tau)}{n}. \tag{9}$$

Its MCF is estimated by (4) and (5) with t_k and $s(t_k)$ being replaced by t_k^* and n, respectively. The MCF obtained in such a way is called as the robust MCF estimator. Its robustness results from the fact that the increment of the MCF is $1/n$, which is smaller than the MCF increment of the Nelson estimator, $1/s(t_k)$.

Based on the robust MCF, the fleet MTBF can be estimated as

$$\mu_{FJ} = \frac{\tau^*}{M(\tau^*)}. \tag{10}$$

It is noted that $M(\tau^*) = N/n$. Using this relation and (9) to (10) yields

$$\mu_{FJ} = \frac{T_3(\tau)}{N} = \mu_F. \tag{11}$$

That is, (11) is consistent with (3) and hence the appropriateness of the robust MCF estimator is confirmed.

2.3. Equivalent acceleration factors

[14] estimate the MTBF of the ith unit as

$$\mu_i = \frac{\tau_i}{n_i}. \tag{12}$$

The unit that has the largest failure number is selected as the reference unit, whose MTBF and censoring time are denoted as μ_R and τ_R, respectively. The EAF of the ith unit is defined as

$$\alpha_i = \frac{\mu_i}{\mu_R}. \tag{13}$$

Clearly, α_i can be viewed as a normalized MTBF, which is dimensionless.

Let μ_α and σ_α denote the mean and standard deviation of all the EAF values, respectively. The fleet heterogeneity is measured by

$$\rho_\alpha = \frac{\mu_\alpha}{\sigma_\alpha}. \tag{14}$$

[14] suggest 1.96 as the critical value of ρ_α. If $\rho_\alpha > 1.96$, the fleet is thought to be homogeneous; otherwise, heterogeneous.

Generally, the MTBF estimated at different time is different and it is negatively correlated with unit age. It is noted that μ_i is estimated at $\tau_i \le \tau$ and μ_R is estimated at τ_R, which is generally different from τ_i. That is, μ_i and μ_R is estimated at different time points. When τ_i and τ_R are significantly different, the EAF estimated by (13) is unreliable. To improve, μ_i and μ_R should be estimated at the same time point τ_i. In this case, (13) is revised as

$$\alpha_i = \frac{\mu_i}{\mu_R(\tau_i)} \tag{15}$$

where $\mu_R(\tau_i)$ is the MTBF of the reference unit estimated at τ_i by interpolation or extrapolation. To facilitate the evaluation of $\mu_R(\tau_i)$ and obtain a more robust estimate of $\mu_R(\tau_i)$, it is necessary to redefine the reference unit. This will yield an improved EAF evaluation method.

2.4. Improved EAF evaluation method

We define a virtual unit as the reference unit, its failure process equals to the "average process" of the fleet, and its MCF is given by the Nelson estimator of the fleet MCF. In this case, we have

$$\mu_R(\tau_i) = \frac{\tau_i}{M(\tau_i)}, \mu_i = \frac{\tau_i}{M_i(\tau_i)} \tag{16}$$

where $M_i(.)$ is the MCF of the ith unit. If $\tau_i > t_{n_i}$, $M_i(\tau_i) = n_i$; if $\tau_i = t_{n_i}$, $M_i(\tau_i) = n_i - 0.5$. Using (16) to (15) yields

$$\alpha_i = \frac{M(\tau_i)}{M_i(\tau_i)}. \tag{17}$$

Assume that the EAF estimated from (17) is approximately a constant. That is, $\alpha_i(\tau_i) \approx \alpha_i(\tau)$. Under this assumption, the MTBFs of the units evaluated at τ can be estimated as

$$\mu_i = \alpha_i \mu_F \tag{18}$$

where μ_F is given by (3). Thus, *the MTBFs of all the units are evaluated at a common time* τ.

It is noted that (18) is the MTBF of the i th unit evaluated at τ while (12) is the MTBF of the i th unit evaluated at τ_i. The two estimates are usually different and their relative error is given by

$$\varepsilon_i = |1 - \frac{\mu_i(\tau_i)}{\mu_i(\tau)}|. \tag{19}$$

When τ_i is much smaller than τ, ε_i may be large, as to be illustrated by Figures 1 and 3 in Section 4.

3. Comparison of the EAF-based approach with frailty model

Traditionally, the heterogeneity of repairable systems in a fleet is modeled by a frailty model [11-13, 31]. In this section, we compare the EAF-based approach with the frailty model and illustrate the conclusions obtained from comparison analysis using a real-world example.

3.1. Frailty model

Consider the failure point processes given by (1). In the context of frailty analysis, each system in the fleet has its own failure intensity; and the effect of the heterogeneity on the failure process of a system is described a non-negative random variable Z called the frailty. The frailty model is defined in terms of the intensity function of a failure process, given by

$$m(t|Z) = Zm_0(t), \tag{20}$$

where $m_0(t)$ is the baseline failure intensity function. Equation (20) is actually a proportional intensity model [32]. According to (20), the intensity function of the ith unit can be written as

$$m_i(t) = z_i m_0(t) \tag{21}$$

where z_i is the frailty of the ith unit.

To specify a frailty model, one needs to specify the type of the repair process (i.e., minimal, imperfect or perfect repair) and the distribution of the frailty. For a minimal repair process, the non-homogeneous Poisson process with a power-law MCF is assumed. The power-law MCF is given by

$$M(t) = (t/\eta)^{\beta}. \tag{22}$$

The frailty is usually assumed to follow the gamma distribution with scale parameter θ and shape parameter $1/\theta$. The expectation of this distribution is one and the variance is θ. A large value of θ indicates that there exists heterogeneity between the systems. This frailty model has three parameters (i.e., β, η; θ) to be estimated. For the data given by (1), the parameters can be estimated using the maximum likelihood estimation method. Since z_i is unobserved, the ordinary likelihood function has to be replaced by the marginal likelihood obtained by unconditioning with respect to z_i.

Once the parameters are specified, the frailty z_i of the ith unit can be estimated. For the above-mentioned frailty model, the frailty of the ith unit is estimated by

$$z_i = (\tfrac{1}{\theta} + n_i)/[\tfrac{1}{\theta} + M(\tau_i)]. \tag{23}$$

3.2. Comparison of the EAF-based approach with the frailty model

We compare the EAF-based approach with the frailty model from four aspects: definitions of EAF and frailty, evaluation methods of heterogeneity, determination of fleet heterogeneity and interpretations of EAF and frailty.

Definition of EAF and frailty: Both the EAF and frailty are defined relative to an "average" system; the EAF is defined in terms of the MTBF, and the frailty is defined in terms of intensity function. From (23), the frailty is asymptotically reciprocally proportional to the EAF as $\theta \to \infty$. These imply that the two models are somehow similar and closely related.

Evaluation of heterogeneity: Parameter θ of the frailty model describes the heterogeneity and is estimated using a parametric approach. The approach needs to make the assumptions about the types of repair process and frailty distribution, and the parameter estimation needs to use an unconditional approach. These make this approach complex and the results unreliable due to possible misspecification. On the other hand, the EAF-based approach does not make any assumptions since it is essentially a non-parametric approach. Therefore, it is simpler and more reliable than the frailty model.

Determination of fleet heterogeneity: The EAF-based approach defines a critical value of ρ_α to objectively determine whether the heterogeneity exists while the frailty-based approach does not define such a critical value so that the determination of fleet heterogeneity is subjective to some extent.

Interpretation of EAF and frailty: The EAF of a system is the ratio of two statistics (i.e., system MTBF and fleet MTBF) while the frailty is the ratio of two functions (i.e., system intensity and baseline intensity). Thus, the former can be viewed as being "observable" since the MTBF can be directly computed from the data while the latter is unobserved since it can be only estimated based on the data and assumed models.

3.3. A robust estimate for frailty of a system

It is noted that the frailty given by (23) is highly sensitive to the value of θ: when $\theta \to 0$, $z_i \to 1$; and when $\theta \to \infty$, $z_i \to \frac{n_i}{M(\tau)} = \frac{1}{\alpha_i}$. To make z_i robust, we define a non-parametric estimate of the frailty. Consider the single censoring case with $\tau_i = \tau, (1 \le i \le n)$. Integrating the two sides of (21) from zero to τ, we have

$$LHS = n_i,\ RHS = z_i M(\tau). \tag{24}$$

Letting $LHS = RHS$ yields

$$z_i = \frac{n_i}{M(\tau)} = \frac{1}{\alpha_i}. \tag{25}$$

This estimate is independent of θ and can be interpreted as the ratio of the system's failure numbers and expected failure number of the fleet in a common observation window $(0, \tau)$. The frailty given by (25) is reciprocally proportional to the EAF. According to this definition, the frailty becomes "observable". The definition can be extended to the multiple censoring case: *the frailty of system i is the ratio of the system's failure numbers and expected failure number of the fleet evaluated at τ_i.*

3.4. Example 1

In this subsection, we illustrate the frailty-based and EAF-based approaches by a real-world example.

3.4.1. Data

The example comes from [31] and deals with the failure processes of 9 sugarcane harvesters, observed in a common period of 200 days. [31] Present the failure time data of the systems with a figure, from which we extract the data shown in the first two rows of Table 1.

Table 1: Failure numbers of sugarcane harvesters in 200 days

i	1	2	3	4	5	6	7	8	9	δ
n_i	11	14	14	15	19	14	11	13	16	0.1752
z_i, (23)	0.9989	0.9991	0.9991	0.9991	0.9994	0.9991	0.9989	0.9990	0.9992	0.0002
z_i, (25)	0.7795	0.9921	0.9921	1.063	1.346	0.9921	0.7795	0.9213	1.134	0.1799
μ_i	18.18	14.29	14.29	13.33	10.53	14.29	18.18	15.38	12.5	0.1704
α_i	1.283	1.008	1.008	0.941	0.743	1.008	1.283	1.085	0.882	0.1704

3.4.2. Results of the frailty-based approach

In [31], the authors estimated assume that the repair is perfect, the time to failure follows the Weibull distribution with shape parameter β and scale parameter η, and the frailty follows the gamma distribution with parameter θ. The maximum likelihood estimates (MLEs) of the parameters are $(\beta, \eta;\ \theta) = (1.32,\ 15.02;\ 0.56\times 10^{-4})$. Since the estimate of θ is nearly equal to zero, the unobserved heterogeneity among the systems is deemed to be non-existent. The frailties estimated from (23) are shown in the third row of Table 1 and the frailties estimated from (25) are shown in the fourth row. The last column shows the coefficient of variation (δ) of the nine values of each row. It is expected that $\delta(z_i)$ should be close to $\delta(n_i)$. However, (23) does not meet this expectation while (25) meets the expectation, and hence its appropriateness is confirmed.

3.4.3. Results of the proposed approach

Since the observation windows of the systems are the same, it is easy to evaluate the MCF and MTBF of the fleet, which are $M(200) = 14.11$ and $\mu_F = 14.17$ days, respectively. It is also easy to calculate the MTBFs and EAFs of individual systems using the following relations

$$\mu_i = 200/n_i,\ \alpha_i = M(200)/n_i \tag{26}$$

The results are shown in the last two rows of Table 1. The average and standard deviation of the EAFs are μ_α =1.0267 and σ_α =0.1749, respectively; and their ratio is $\rho_\alpha = 5.870$, which is much larger than 1.96. Therefore, the fleet is homogeneous. This conclusion is consistent with the one of [31] though $\delta(z_i)$ obtained from (23) is significantly different from $\delta(z_i)$ obtained from (25).

4. Further illustrations

In this section, we further illustrate the proposed approach and its superiority using two examples, which have been analyzed by [14].

4.1. Example 2

4.1.1. Data

Consider a fleet of 10 units, and each has a power-law MCF given by (22) with parameters $\beta = 1.5$ and $\eta = 100$. A failure is restored by minimal repair; and the censoring time of each unit is randomly generated from the uniform distribution over (0, 500) so that $\tau_i > t_{n_i}$. The randomly generated data are shown in the upper part of Table 2 and the censoring times of units are shown in the last row. From the table yields four important statistics:

$$N = 46,\ \tau = 470.9,\ T_3(\tau) = 2601.7,\ \mu_F = 56.56. \tag{27}$$

Table 2: Failure processes of units

Unit	1	2	3	4	5	6	7	8	9	10
t_1	245	131		17	105		19	70	108	32
t_2	293	194		51	209		49	174	188	103
t_3	353	277		56	235			195	235	
t_4	356	278		188	255			215		
t_5	378	279		206	262			271		
t_6	401	324		216	270			356		
t_7	424			231	281					
t_8	466			249	331					
t_9				251						
t_{10}				338						
t_{11}				372						
τ_i	470.9	336.1	50.5	441.3	348.3	111.6	92.8	374.5	239.4	136.3

4.1.2. Evaluation of MTBF and EAF

The information that the approach of in [14], the authors estimated requires for estimating the MTBFs and EAFs of units are shown in the first three columns of Table 3; and the estimated MTBFs and EAFs are shown in the 4th and 5th columns. As can be seen, using the approach of [14] cannot estimate the MTBFs and EAFs of Units 3 and 6 due to $n_3 = n_6 = 0$.

Although no failure occurs for Units 3 and 6, these two units can and should be included in the computation of MCF since their censoring information has an influence on $s(t_k)$. The results obtained from the proposed approach are shown in the 6th to 8th columns. Different from the approach of [14], which first estimates μ_i (at τ_i) and then estimates α_i, the improved method first estimates α_i (at τ_i) and then estimates μ_i (at τ).

Table 3: MTBFs and EAFs of units obtained from two methods

Unit	Jiang et al. [14]				Improved approach			ε_i
	τ_i	n_i	μ_i (12)	α_i (13)	$M(\tau_i)$	α_i (17)	μ_i (18)	
1	470.9	8	58.86	1.467	10.13	1.267	71.65	0.1785
2	336.1	6	56.02	1.396	6.051	1.009	57.04	0.0180
3	50.50	0			0.4000	(0.7315)	(41.37)	
4	441.3	11	40.12	1.000	9.135	0.8304	46.97	0.1458
5	348.3	8	43.54	1.085	6.301	0.7876	44.55	0.0227
6	111.6	0			1.108	(2.027)	(114.6)	
7	92.80	2	46.40	1.157	0.7333	0.3667	20.74	1.237
8	374.5	6	62.42	1.556	7.635	1.272	71.97	0.1327
9	239.4	3	79.80	1.989	3.251	1.084	61.29	0.3019
10	136.3	2	68.15	1.699	1.251	0.6256	35.38	0.9261

Since the reference process is an "average process", it is possible to estimate the MTBFs and EAFs of Units 3 and 6 based on the following two assumptions:

(a) The average of EAFs of all the units equals to 1, and

(b) From (17) we can assume that $\alpha_i = cM(\tau_i)$ for $i = 3$ and 6.

The value of c can be estimated from the first assumption, which is 1.829. From the second assumption yields $\alpha_3 = 0.7315$ and $\alpha_6 = 2.027$. Further, from (18) yields $\mu_3 = 41.37$ and $\mu_6 = 114.6$. These estimates are bracketed in Table 3.

The last column shows the relative errors between the MTBFs estimated from (12) and (18). To examine the influence of censoring time τ_i (or failure number n_i) on the accuracy of MTBF estimate, we examine the plot of ε_i vs.

n_i shown in Figure 1. As seen from the figure, ε_i quickly increases as n_i decreases (or $1/n_i$ increases), implying that the MTBF estimates obtained from (12) and (18) are nearly the same when n_i is large. Since the estimate obtained from (12) is accurate when n_i is large, this actually validates the appropriateness of (18).

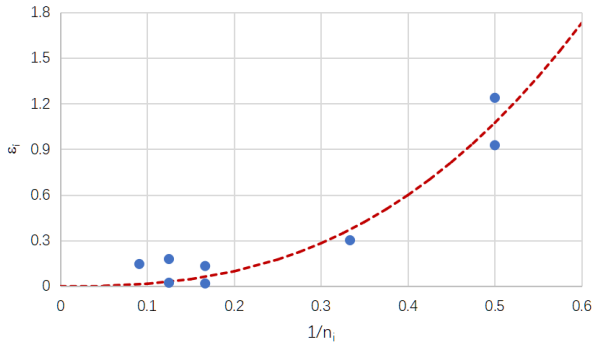


Figure 1: Plot of ε_i *vs.* n_i for Example 2

4.1.3. *Estimates of fleet MTBF*

As mentioned in Section 2, the fleet MTBF can be estimated by a TTT-based approach given by (3) and by an MCF-based approach given by (6). We illustrate that they are different for multiple censoring case.

From (27) and Table 3 we have $\tau = \tau_1 = 470.9$ and $M(\tau) = M(\tau_1) = 10.13$. Applying these to (6) yields $\mu_{FN} = 46.46$. On the other hand, from (3) we have $\mu_F = 56.56$. This implies that the MCF-based approach underestimates the fleet MTBF with a relative error of 17.85%. This also indirectly illustrates that the Nelson MCF estimator can be inaccurate for the multiple censoring case, and its accuracy can be measured by the relative error given by

$$\varepsilon_F = |1 - \frac{\mu_{FN}}{\mu_F}|. \tag{28}$$

The larger the relative error is, the less accurate the Nelson MCF estimator is.

4.1.4. *Heterogeneity evaluation*

The computational results of the heterogeneity are shown in Table 4. The results obtained from the approach of [14] are shown in the second row, the results obtained from the improved method with the EAFs of Units 3 and 6 being excluded are shown in the third row, and the results obtained from the improved method with the EAFs of Units 3 and 6 being included are shown in the last row. As seen, the values of ρ_α for these three cases are quite different, but they all are larger than 1.96, implying that the fleet is homogeneous.

Table 4: Fleet heterogeneity measures obtained from different methods

Approach	Units 3 and 6	μ_α	σ_α	ρ_α
[14]	Excluded	1.419	0.3340	4.248
This paper	Excluded	0.9052	0.3153	2.871
	Included	1.000	0.4587	2.180

The difference in ρ_α can be explained by a large ε_F. Since Units 7, 9 and 10 only have 2, 3 and 2 failures, respectively, their MTBFs are considerably overestimated by (12). This reduces the dispersion of the MTBFs and EAFs and leads to a large value of ρ_α. This confirms the fact that the units' EAF estimates obtained from the proposed approach are more robust than those obtained from the approach of [14].

4.1.5. *Distribution of EAF*

The frailty is usually assumed to follow the gamma distribution with mean one. It is possible to check whether this assumption holds for the EAF. Fitting the EAFs shown in the 7th column of Table 3 to four typical 2-parameter distributions: gamma distribution with parameters u and v, Weibull distribution with parameters β and η, normal distribution with parameters μ and σ, and lognormal distribution with parameters μ_l and σ_l. The MLEs of the parameters are shown in Table 5, where $\theta_1 = u, \beta, \mu$ or μ_l; $\theta_2 =, \eta, \sigma, v$ or σ_l; and ln (L) is the log-likelihood value. As seen, the gamma distribution provides the best fit in terms of ln (L).

Table 5: MLEs of distributional parameters

	Example 2				Example 3			
	Gamma	Weibull	Normal	Lognormal	Gamma	Weibull	Normal	Lognormal
θ_1	5.498	2.434	1.000	-0.0936	4.463	2.005	1.229	0.0896
θ_2	0.1819	1.130	0.4352	0.4413	0.2753	1.397	0.6631	0.4580
$ln(L)$	-5.034	-5.423	-5.871	-5.074	-10.37	-11.50	-13.11	-9.459

The probability plot of a distribution is often used to check the appropriateness of the distribution. Let α_j $(1 \le j \le n)$ denote the ordered EAFs sorted in ascending order, $p_j = (j - 0.3)/(n + 0.4)$, and $q_j = F^{-1}(p_j;\, u, 1)$ denote the inverse function of the gamma distribution with parameters u and 1. The gamma probability plot is given by

$$\alpha_j = vq_j. \tag{29}$$

Figure 2 shows the plot of α_j vs. q_j for $u = 5.498$. Regression yields $v = 0.1879$, which is close to the MLE of v with a relative error of 3.29%. This confirms the appropriateness of the gamma distribution for fitting the EAF data.

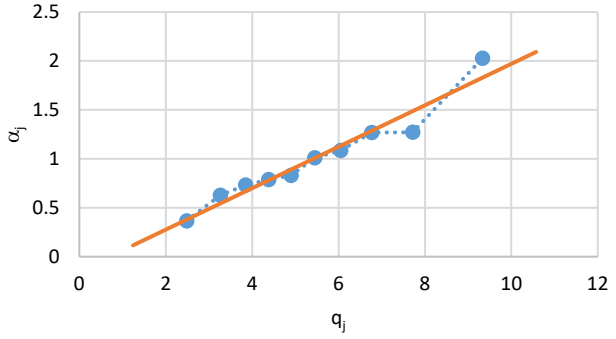


Figure 2: Plot of α_j vs. q_j for Example 2

4.2. Example 3

4.2.1. Data

The data shown in Table 6 come from Proschan [33] and deal with the inter-failure times (in days) of 13 air conditioning systems before overhauls. For this example, the censoring time meets the relation of $\tau_i = t_{n_i}$. From the table, we have

$$N = 192, \tau = 2074, T_3(\tau) = 18088, \mu_F = 94.21. \quad (30)$$

Table 6: Inter-failure times of air conditioning systems

$j\backslash i$	1	2	3	4	5	6	7	8	9	10	11	12	13
1	194	413	90	74	55	23	97	50	359	50	130	487	102
2	15	14	10	57	320	261	51	44	9	254	493	18	209
3	41	58	60	48	56	87	11	102	12	5		100	14
4	29	37	186	29	104	7	4	72	270	283		7	57
5	33	100	61	502	220	120	141	22	603	35		98	54
6	181	65	49	12	239	14	18	39	3	12		5	32
7		9	14	70	47	62	142	3	104			85	67
8		169	24	21	246	47	68	15	2			91	59
9		447	56	29	176	225	77	197	438			43	134
10		184	20	386	182	71	80	188				230	152
11		36	79	59	33	246	1	79				3	2
12		201	84	27		21	16	88				130	14
13		118	44			42	106	46					230
14			59			20	206	5					66
15			29			5	82	5					61
16			118			12	54	36					34
17			25			120	31	22					
18			156			11	216	139					
19			310			3	46	210					
20			76			14	111	97					
21			26			71	39	30					
22			44			11	63	23					
23			23			14	18	13					
24			62			11	191	14					
25						16	18						
26						90	163						
27						1	24						
28						16							
29						52							
30						95							

4.2.2. Estimation of MTBF and EAF

The results obtained from the approach of [14] are outlined in the 2nd to the 5th column of Table 7. The Nelson MCF estimators at unit censoring times are shown in the 6th column and the EAFs and MTBFs of units obtained from the improved methods are shown in the 7th and 8th columns, respectively. It is noted that $\tau = \tau_7 = 2074$ and $M(\tau) = M(\tau_7) = 23.74$. Applying these to (6) yields $\mu_{FN} = 87.38$, which is slightly smaller than the fleet MTBF shown in (30) with a relative error of 7.25%. An underestimate of μ_{FN} implies an overestimate of $M(\tau)$. Compared with Example 2, ε_F is relatively small.

Figure 3 shows the plot of ε_i vs. n_i. As expected, ε_i increases as n_i decreases. The average and maximum of ε_i's are 0.0683 and 0.2258, respectively, implying that the MTBF estimates obtained from (12) and (18) are fairly close to each other for the current example.

Table 7: MTBFs and EAFs of units obtained from two methods for Example 3

Unit	[14]				Improved approach			ε_i
	τ_i	n_i	μ_i (12)	α_i (13)	$M(\tau_i)$	α_i (17)	μ_i (18)	
1	493	6	82.16	1.378	4.269	0.7115	67.03	0.2258
2	1851	13	142.4	2.389	19.99	1.537	144.8	0.0167
3	1705	24	71.04	1.191	18.80	0.7835	73.81	0.0375
4	1314	12	109.5	1.837	12.95	1.080	101.7	0.0767
5	1678	11	152.5	2.559	18.22	1.656	156.0	0.0221
6	1788	30	59.60	1.000	19.28	0.6426	60.54	0.0155
7	2074	27	76.81	1.288	23.74	0.8791	82.82	0.0725
8	1539	24	64.12	1.075	16.23	0.6763	63.71	0.0065
9	1800	9	200.0	3.355	19.57	2.174	204.8	0.0234
10	639	6	106.5	1.786	6.111	1.018	95.95	0.1100
11	623	2	311.5	5.226	5.933	2.966	279.5	0.1145
12	1297	12	108.1	1.813	12.71	1.059	99.80	0.0830
13	1287	16	80.43	1.349	12.61	0.7879	74.23	0.0836

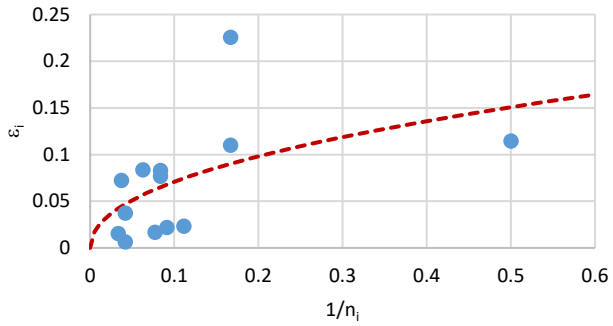


Figure 3: Plot of ε_i vs. n_i for Example 3

4.2.3. Heterogeneity evaluation

The fleet heterogeneity measure obtained from [14] is $\rho_\alpha = 1.715$, implying that the fleet is heterogeneous. The value of ρ_α obtained from the improved method is 1.780, which is very close to the above estimate. The similarity of the results obtained from the two approaches can be explained by the average failure number of units (i.e., N/n), which is 14.8. For a dataset with a large value of N/n, both the method of [14] and the improved method will give similar results and draw similar conclusions; otherwise, the results and conclusions are probably different.

4.2.4. Distribution of EAF

Fitting the EAFs shown in the 7th column of Table 7 to the four 2-parameter distributions yields the MLEs of the parameters shown in the RHS of Table 5. In terms of the log-likelihood value, the lognormal distribution provides the best fit to the data. To further validate, we examine the lognormal probability plot. Let $q_j = F^{-1}(p_j; 0, 1)$ denote the inverse function of the standard normal distribution. The lognormal probability plot is given by

$$q_j = \frac{ln(\alpha_j) - \mu_l}{\sigma_l}. \quad (31)$$

Figure 4 shows the plot of q_j vs. $ln\ (\alpha_j)$. As can be seen, the plot is concave, implying that the 2-parameter lognormal distribution is not appropriate for fitting the EAF data.

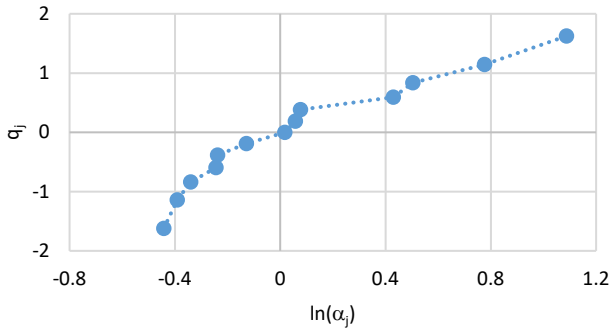


Figure 4: Plot of q_j vs. ln (α_j) for EAF data of Example 3

Figure 4 indicates that urther analysis shows that the 3-parameter lognormal distribution with a location parameter γ can be appropriate for fitting the data. This is validated by the probability plot of the 3-parameter lognormal distribution shown in Figure 5. The least square estimates of the parameters are $\gamma = 0.5934$, $\mu_l = -1.016$, $\sigma_l = 1.259$, and the corresponding log-likelihood value is $\ln(L) = -6.854$, which is much larger than the log-likelihood value of the 2-parameter lognormal distribution. From the fitted model yields $\rho_\alpha = 0.8843$, and hence the heterogeneity of the fleet is confirmed.

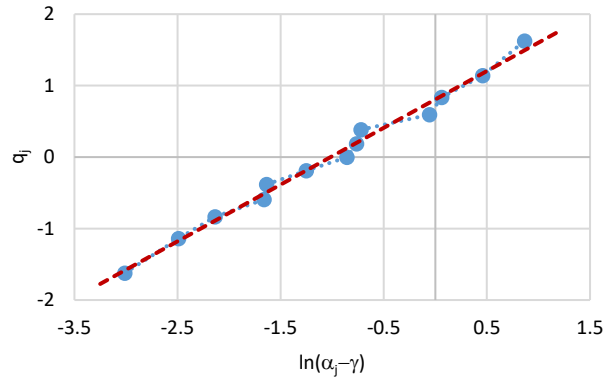


Figure 5: Probability plot of the 3-parameter lognormal distribution of the EAF for Example 3

5. Conclusions

In this paper, we have identified the problems of the heterogeneity evaluation method presented in [14] and proposed an improved method, whose main advantages are:

(a) it estimates the MTBFs of units at a common time point so that the results are much more reliable than those obtained from the method given in [14], and

(b) the MTBFs of zero-failure units can be inferred under two reasonable assumptions. This is different from those approaches that exclude the units with an inadequate amount of failure data from the analysis [16].

These have been illustrated by three examples.

The main conclusions and/or findings of the paper have been:

- The MTBFs and EAFs of units estimated from the method [14] and the improved method are fairly consistent if the average failure number of units is large and can be considerably different otherwise.
- The fleet MTBF estimates obtained from the Nelson MCF estimator and TTT approach are different for multiple censoring data.
- The frailty of a system can be defined as the ratio of the system's failure numbers and expected failure number of the fleet evaluated in a common observation window. Such defined frailty is reciprocally proportional to the EAF.
- A simple distribution such as the 2-parameter gamma or lognormal distribution may be inappropriate for describing the EAF or frailty in the presence of heterogeneity.

In this paper, we have confined to the fleet composed by similar repairable systems rather than multi-component repairable systems and stressed MTBFs of units rather than their trends. The MTBF estimates of the units are based on the constant EAF assumption, whose reasonability requires further validation. The data considered in this paper only contain failure time information without other information such as failure causes, failure modes, maintenance types and censoring types. Analysis for the failure process with such information is a challenging issue and needs further research. Finally, a topic for future study is to develop a cluster analysis method for classifying the units of a heterogeneous fleet (e.g., automated guided vehicles [34]) into several homogeneous groups based on the EAFs of the units.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the referees for their helpful comments and suggestions which have greatly enhanced the clarity of the paper. This work was supported by the 'Pioneer and Leading Goose + X' R&D Program of Zhejiang Province, China (2024SJCZX0037) and Wenzhou Major Science and Technology Innovation Projects (ZG2023019).

References

[1] Ascher H, Feingold H. *Repairable Systems – Modeling, Inference, Misconceptions and their Causes*. Marcel Dekker, New York, 1984.

[2] Lindqvist BH. "Maintenance of repairable systems," in *Complex system maintenance handbook*, pp. 235-261, 2008, doi: 10.1007/978-1-84800-011-7_10

[3] Lindqvist BH. "On the statistical modeling and analysis of repairable systems," *Statistical Science*, vol. 21, no. 4, pp. 532-551, 2007, doi: 10.1214/088342306000000448.

[4] Krivtsov VV. "Practical extensions to NHPP application in repairable system reliability analysis," *Reliability Engineering and System Safety*, vol. 92, no. 5, pp. 560-562, 2007, doi: 10.1016/j.ress.2006.05.002.

[5] Jiang R, Huang C. "Failure patterns of repairable systems and a flexible intensity function model," *International Journal of Reliability and Applications*, vol. 13, no. 2, pp. 81-90, 2012, doi: koreascience.kr/article/JAKO201217752421606.

[6] R. Jiang, Y. Guo. and strong, "Estimating Failure Intensity of a Repairable System to decide on its Preventive Maintenance or Retirement," *International journal of performability engineering*, vol. 10, no. 6, pp. 577-588, 2014, doi: 10.23940/ijpe.14.6.p577.mag.

[7] Nelson W. "Graphical analysis of system repair data," *Journal of Quality Technology*, vol. 20, pp. 24-35, 1988, doi: 10.1080/00224065.1988.11979080.

[8] Nelson W. *Recurrent Events Data Analysis for Product Repairs, Disease Recurrences, and Other Applications*, Society for Industrial and Applied Mathematics (SIAM): Philadelphia, 2003.

[9] Asfaw ZG, Lindqvist BH. "Unobserved heterogeneity in the power law nonhomogeneous Poisson process," *Reliability Engineering and System Safety*, vol. 134, pp. 59-65, 2015, doi: 10.1016/j.ress.2014.10.005

[10] Cui D., Sun Q., Xie M., "Robust statistical modeling of heterogeneity for repairable systems using multivariate gaussian convolution processes," *IEEE Transactions on Reliability*, vol. 72, no. 4, pp. 1493-1506, 2023, doi: 10.1109/TR.2023.3235889.

[11] Liu X., Vatn J., Dijoux Y., Toftaker H., "Unobserved heterogeneity in stable imperfect repair models," *Reliability Engineering and System Safety*, vol. 203, pp. 107039, 2020, doi: 10.1016/j.ress.2020.107039.

[12] Zaki R, Barabadi A, Barabady J, Qarahasanlou AN. "Observed and unobserved heterogeneity in failure data analysis," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, vol. 236, no. 1, pp. 194-207, 2022, doi: 10.1177/1748006X211022538

[13] Brown B., Liu B., McIntyre S., Revie M., "Reliability evaluation of repairable systems considering component heterogeneity using frailty model," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, SAGE journal*, vol. 237, no. 4, pp. 654-670, 2023, doi: 10.1177/1748006X221109341.

[14] Jiang R., Li F., Xue W., Cao Y., et al., "A robust mean cumulative function estimator and its application to overhaul time optimization for a fleet of heterogeneous repairable systems," *Reliability Engineering and System Safety*, vol. 236, pp. 109265, 2023, doi: 10.1016/j.ress.2023.109265.

[15] Garmabaki, AHS, Ahmadi A, Block J, Pham H, Kumar U. "A reliability decision framework for multiple repairable units," *Reliability Engineering & System Safety*, vol. 150, pp. 78-88, 2016, http://dx.doi.org/10.1016/j.ress.2016.01.020

[16] Garmabaki AHS, Ahmadi A, Mahmood YA, Barabadi A. "Reliability modelling of multiple repairable units," *Quality and Reliability Engineering International*, vol. 32, no. 7, pp. 2329-2343, 2016, doi: 10.1002/qre.1938

[17] Weckman GR, Shell RL, Marvel JL. "Modeling the reliability of repairable systems in the aviation industry," *Computers and Industrial Engineering*, vol. 40, pp. 51-63, 2001, doi: 10.1016/S0360-8352(00)00063-2.

[18] Giorgio M, Guida M, Pulcini G. "Repairable system analysis in presence of covariates and random effects," *Reliability Engineering and System Safety*, vol. 131, pp. 271-281, 2014, doi: 10.1016/j.ress.2014.04.009.

[19] Meeker WQ, Hong Y. "Reliability meets big data: opportunities and challenges," *Quality Engineering*, vol. 26, pp. 102-116, 2014, doi: 10.1080/08982112.2014.846119.

[20] Navas MA, Sancho C., Carpio J. "Reliability analysis in railway repairable systems," *International Journal of Quality & Reliability Management*, vol. 34, no. 8, pp. 1373-1398, 2017, doi: 10.1108/IJQRM-06-2016-0087.

[21] Hong Y, Zhang M, Meeker WQ. "Big data and reliability applications: the complexity dimension," *Journal of Quality Technology*, vol. 50, no. 2, pp. 135-149, 2018, doi: 10.1080/00224065.2018.1438007.

[22] Peng W, Shen L, Shen Y, Sun Q, "Reliability analysis of repairable systems with recurrent misuse-induced failures and normal-operation failures," *Reliability Engineering and System Safety*, vol. 171, pp. 87-98, 2018, doi: 10.1016/j.ress.2017.11.016.

[23] Si W, Love E, Yang Q. "Two-state optimal maintenance planning of repairable systems with covariate effects," *Computers & Operations Research*, vol. 92, pp. 17-25, 2018, doi: 10.1016/j.cor.2017.11.007.

[24] Liu X, Pan R. "Analysis of large heterogeneous repairable system reliability data with static system attributes and dynamic sensor measurement in big data environment," *Technometrics*, vol. 62, no. 2, pp. 206-222, 2020, doi: 10.1080/00401706.2019.1609584.

[25] Sharma G, Rai RN. "Failure modes based censored data analysis for repairable systems and its industrial perspective," *Computers & Industrial Engineering*, vol. 158, pp. 107439, 2021, https://doi.org/10.1016/j.cie.2021.107439

[26] Jiang R, Li F, Xue W, Lin L, Li X, Zhang K. "Identification and treatment of extreme inter-failure times from a fleet of repairable systems,". In G. Abdul-Nour et al. (eds.), *17th WCEAM Proceedings*, Lecture Notes in Mechanical Engineering, pp. 417-431, 2023, https://doi.org/10.1007/978-3-031-59042-9_34

[27] Jiang R, Xue W, Cao Y. "Analysis for influence of maintenance and manufacturing quality on reliability of repairable systems," In *Advances in Reliability and Maintainability Methods and Engineering Applications*, pp. 385-403. Cham: Springer Nature Switzerland, 2023, https://doi.org/10.1007/978-3-031-28859-3_15

[28] Jiang R. "Overhaul decision of repairable systems based on the power-law model fitted by a weighted parameter estimation method," *In Asset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies*, Chapter 29, pp. 277-286, Springer, 2018. doi: 10.1007/978-3-319-95711-1_28.

[29] Block J, Ahmadi A, Tyrberg T, Kumar U. "Fleet-level reliability analysis of repairable units: a non-parametric approach using the mean cumulative function," *International Journal of Performability Engineering*, vol. 9, no. 3, pp. 333-344, 2013.

[30] Block J, Ahmadi A, Tyrberg T, Kumar U. "Fleet-level reliability of multiple repairable units: a parametric approach using the power law process," *International Journal of Performability Engineering*, vol. 10, no. 3, pp. 239-250, 2014, doi: 10.23940/ijpe.14.3.p239.mag

[31] Brito ÉS, Tomazella VLD, Ferreira PH. "Statistical modeling and reliability analysis of multiple repairable systems with dependent failure times under perfect repair," Reliability Engineering and System Safety, vol. 222, pp. 108375, 2022, doi: 10.1016/j.ress.2022.108375.

[32] Percy DF, Alkali BM. "Generalised proportional intensities models for repairable systems," *IMA Journal of Management Mathematics*,

vol. 17, pp. 171-185, 2006, doi: 10.1093/imaman/dpi034.

[33] Proschan F. "Theoretical explanation of observed decreasing failu re rate," *Technometrics*, vol. 5, no. 3, pp. 375-383, 1963, doi:10.1080/ 00401706.1963.10490105.

[34] Rashidi H, Matinfar F, Parand FA. "Automated guided vehicles-a review on applications, problem modeling and solutions," *Intern ational Journal of Transportation Engineering*, vol. 8, no. 3, pp. 261-2 78, 2021, doi: 10.22119/IJTE.2021.246669.1531.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/)



RENYAN JIANG received his B.S. and M.S. degrees from Wuhan University of Technology, China, in 1982 and 1985, respectively and his Ph.D. in 1996 from University of Queensland, Australia. His research interests are in various aspects of quality, reliability and maintenance. He is the author or co-author of five reliability/maintenance related books and has published more than 250 papers.



XIA XU has done his bachelor's degree from Wuhan University of Technology in 2005. He is the leading craftsman at Penta Laser (Zhejiang) Co., Ltd. He invented 15 patents, published 9 papers, and won the first prize of Zhejiang Machinery Industry Sci. & Tech.



KUNPENG ZHANG received his B.S. (2013) degree from Nanyang Institute of Technology, M.S. (2016) and Ph.D. (2021) degrees from Xiamen University. His research interests are micro/nano 3D printing, microfluidic and the reliability engineering and system safety.



YU CAO received his B.S. (2003) degree from Harbin Institute of Technology (HIT), M.S. (2006) and Ph.D. (2009) degrees from Huazhong University of Science & Technology (HUST). His research interests are laser material processing, Reliability Engineering and System Safety.

Received: 16 September 2024, Revised: 14 October 2024, Accepted: 15 October 2024, Online: 31 October 2024
DOI: https://dx.doi.org/10.55708/js0310007

Enhancing Mental Health Support in Engineering Education with Machine Learning and Eye-Tracking

Yuexin Liu[*,1], **Amir Tofighi Zavareh**[1], **Ben Zoghi**[2]

[1]Department of Engineering Technology and Industrial Distribution, Texas A&M University, College Station, 77843, USA
[2]Lyle School of Engineering, Southern Methodist University, Dallas, 75205, USA
*Corresponding author: Yuexin Liu, Email: Yuexin.liu@tamu.edu

ABSTRACT: Mental health concerns are increasingly prevalent among university students, particularly in engineering programs where academic demands are high. This study builds upon previous work aimed at improving mental health support for engineering students through the use of machine learning (ML) and eye-tracking technology. A framework was developed to monitor mental health by analyzing eye movements and physiological data to provide personalized support based on student behavior. In this extended study, baseline data were analyzed to explore the correlations between emotions and physiological biomarkers. Key findings indicate that emotions such as Anger and Fear are positively correlated with increased physical activity, while Sadness is associated with elevated respiratory rates. A strong positive correlation between Electrodermal Activity (EDA) and Happiness was also identified, indicating physiological markers linked to positive emotional states. Temporal patterns were observed, with heightened emotional tagging occurring more frequently in the evening. These findings deepen the understanding of how emotional states manifest through physiological changes, providing a foundation for enhancing real-time, personalized mental health interventions. The results contribute to a more comprehensive framework for supporting student well-being and academic performance within engineering education.

KEYWORDS: Mental Health, Philosophy of Engineering Education, Data Correlation, Factor Analysis, Machine Learning, Electrodermal Activity

1. Introduction

Mental health concerns have become increasingly prominent among university students, especially within engineering programs, where students are subject to intense academic demands and pressure [1, 2]. Traditionally, engineering education has focused on the development of technical knowledge and skills [3, 4], often placing limited emphasis on the physical and mental well-being of students. As a result, many students in engineering experience high levels of stress, which can negatively impact their academic performance, personal lives, and overall professional growth.

Current methods for assessing mental health, such as surveys and self-reported questionnaires, have limitations in terms of accuracy and bias. These methods may not fully capture the extent of students' mental health challenges, as some may feel reluctant to disclose personal issues due to stigma, fear of judgment, or concerns about potential academic consequences [5]. While initiatives like counseling services, awareness campaigns, and stress management workshops have been introduced to address these issues , there is a pressing need for more objective and effective methods to monitor and improve the mental health of engineering students [6].

Predictive analytics, particularly when applied through machine learning (ML) models, presents a promising solution for addressing mental health challenges. In healthcare, ML techniques have shown potential to transform mental health assessment and intervention by providing objective insights into emotional and psychological states. Building on this, the conference paper Improving Mental Health Support in Engineering Education Using Machine Learning and Eye-Tracking introduced a framework leveraging ML and eye-tracking technology to monitor student well-being. This innovative approach aimed to fill gaps in mental health monitoring by providing data-driven insights and personalized interventions.

This study introduces insights based on data from a 10-week study involving 18 participants. It includes detailed correlations between emotional states—such as Anger, Fear, Happiness, and Sadness—and physiological markers, including physical activity (actigraphy), respiratory rate, and Electrodermal Activity (EDA). Results reveal that increased actigraphy counts are associated with Anger and Fear, while Sadness correlates with elevated respiratory rates. Furthermore, a strong positive relationship between EDA and Happiness is observed. These findings offer a deeper understanding of the physiological responses to emotional states, improving the system's ability to provide real-time, personalized mental health support.

This paper is structured as follows: the next section provides background information on machine learning algorithms and mental health prediction. The methodology section outlines the strategy used for data collection and analysis. The results section presents detailed findings on

the emotional and physiological correlations, followed by a discussion of the potential impact on future mental health interventions. The paper concludes by summarizing the findings and offering recommendations for future research directions.

2. Background

2.1. Machine Learning

Machine learning (ML) refers to the application of statistical and probabilistic methods to build systems that can learn and improve from experience [7, 8]. This capability makes ML a powerful tool for predicting mental health outcomes by analyzing large amounts of complex data, leading to the development of intelligent automated systems that can offer personalized insights. Several algorithms such as support vector machines (SVM), random forests, and artificial neural networks (ANNs) have proven effective in predicting future outcomes and categorizing data. In the healthcare sector, ML is widely applied in various forms, including supervised learning, unsupervised learning, and deep learning. There are also hybrid methods like semi-supervised learning, which combines elements of both supervised and unsupervised learning, as well as reinforcement learning [9].

Supervised learning (SL) is commonly used in healthcare for disease prediction, where the algorithms are trained on a dataset that is pre-labeled with known outcomes. These models are then tested on unseen data to evaluate their predictive performance. On the other hand, unsupervised learning (UL) works without labeled data and is designed to detect patterns and clusters in datasets. Techniques such as k-means clustering, hierarchical clustering, and principal component analysis (PCA) are commonly employed to identify meaningful patterns within the data. In healthcare, UL is valuable for uncovering hidden structures in medical imaging or genetic data, which can help identify subtypes of diseases and facilitate personalized treatment strategies. Additionally, UL is useful for anomaly detection and feature reduction, contributing to more accurate medical diagnoses.

2.2. Deep Learning

Deep learning (DL) [10] is a subset of ML that uses artificial neural networks with multiple layers of nodes to learn intricate representations from raw data. This approach mimics the way the human brain processes information, enabling the discovery of complex relationships within high-dimensional datasets like electronic health records (EHRs). While DL models are highly effective in finding patterns, their multi-layered structure can make it difficult to interpret the decision-making process behind their outputs. Despite this challenge, DL has shown great potential in analyzing vast amounts of healthcare data.

2.3. Reinforcement Learning

Reinforcement learning (RL) is an area of ML where intelligent agents learn to make decisions by interacting with their environment and receiving feedback in the form of rewards or penalties. RL has been successfully applied in areas like robotics [11, 12, 13], gaming [14, 15], education [16] and finance. Recently, RL has gained attention as a promising method for developing personalized mental health interventions [17, 18, 19]. Mental health disorders, being a major source of disability worldwide, often require tailored treatments. RL can create individualized interventions by adapting to patient needs over time based on continuous feedback. One key strength of RL is its ability to work with dynamic and complex data, such as EHRs and wearable device outputs. It is particularly effective in handling incomplete or noisy data, which is frequently encountered in mental health studies. However, implementing RL in mental health care poses challenges related to privacy, transparency, and the interpretability of models. Despite these obstacles, RL holds great promise for creating adaptive, patient-centered treatment plans.

3. Methodology

In recent years, continuous and real-time monitoring technologies have gained significant traction due to the potential to enhance cognitive and behavioral performance while reducing healthcare costs. With the increasing need to monitor mental health, researchers have been investigating different technologies to develop efficient monitoring systems [20, 21]. Among these, eye-tracking technology has shown promise for its ability to track mental health indicators [22]. By applying computational methods to the extensive physiological data collected through eye-tracking, intelligent systems like IntelEye [23] can extract meaningful patterns by analyzing eye movement data such as pupil dilation, fixation points, and blink rates. IntelEye specifically uses the K Nearest Neighbor (KNN) algorithm to classify eye movement patterns into stress levels. The system processes this data by first segmenting the video-watching experience into different scenes, identifying the moments when stress indicators (e.g., dilation of pupils, rapid blinking) are detected. These stress-related signals are then correlated with the specific scenes, allowing IntelEye to detect when stress occurs and to identify the exact content that triggered it. This dual capability, detecting stress and linking it to specific triggers, makes IntelEye a powerful tool for monitoring mental health.

3.1. Study Selection

This section discusses three typical studies conducted over the past decade that explore the advancements in machine learning algorithms for mental health assessment. The focus is on research published from 2015 to 2023, although the review is not systematic and does not cover every possible study meeting the broader criteria. Relevant research was identified using PubMed, ScienceDirect, IEEE, and Google Scholar, with a focus on clinical studies applying machine learning to mental health. Studies involving untested or theoretical ML applications were excluded. The selected studies represent key original research efforts, as outlined in Table 1.

3.2. Review of Selected Studies

Researchers explored a method for automatically assessing depression severity by analyzing facial landmarks

Table 1: Typical ML and Mental Health Studies

Authors	Sample Size	Method	Performance
Anis Kacem et al. [24]	49	SL	84%
Subhagata Chattopadhyay et al. [25]	302	DL	95.5%
Fabian Wahle et al. [26]	28	SL	62%

and 3D head motion using barycentric coordinates and Lie-algebra rotation matrices [24]. Key features were extracted, processed, and encoded using Gaussian Mixture Models (GMM) and Fisher vector encoding. The study, which involved adults with chronic depression, achieved classification accuracy comparable to state-of-the-art deep learning models, while providing interpretable clinical insights.

Building on these computational approaches, a mathematical model was developed to reflect how psychiatrists evaluate depression symptoms [25]. Fourteen symptoms of adult depression were considered, in line with the Diagnostic and Statistical Manual (DSM-IV-TR). Principal Component Analysis (PCA) was used to reduce the number of symptoms to seven key features, which were then input into a hybrid system that combined Mamdani's fuzzy logic controller with a feed-forward multilayer neural network (FFMNN). The system was further refined using a backpropagation neural network (BPNN). This model, validated on 302 real-world depression cases and 50 controls, achieved an average diagnostic accuracy of 95.50

As technological interventions in mental health continue to evolve, [26] investigated the potential of smartphone-based interventions to support individuals with depressive symptoms by collecting data through the Mobile Sensing and Support (MOSS) app. The app gathered context-sensitive sensor data from participants and provided real-time, personalized interventions based on cognitive behavior therapy. Over an eight-week period, participants regularly completed self-reported depression surveys (PHQ-9). For those with clinical depression and high adherence, significant reductions in PHQ-9 scores were observed, indicating the effectiveness of the system in reducing depressive symptoms.

3.3. Results and Future Directions

After reviewing various works in the field, it is clear that eye gaze measures have been employed by multiple models to detect emotions, stress, cognitive load, mental fatigue, and other mental states [27, 28, 29]. This study builds upon these methodologies by developing a system to collect and analyze students' eye movements and physiological responses during remote learning activities, such as attending online lectures. Using machine learning (ML) algorithms, the data will be analyzed to identify patterns and anomalies that indicate changes in mental health, such as stress, anxiety, depression, or distraction. The model will be trained using a combination of eye-tracking data and mental health survey results. The system will then provide real-time feedback to both students and instructors, offering personalized support strategies, such as relaxation exercises or mindfulness training.

The system will be developed and tested in two phases. In the first phase, participants will watch videos designed to elicit both positive and negative emotions while their eye movements are tracked. The gathered data will then be used to train a reinforcement learning (RL) model, which will learn to identify emotional states based on eye movements. In the second phase, the trained model will be integrated into an online platform that provides personalized mental health interventions based on the detected emotional states. This could include relaxation exercises, motivational messages, or other interventions aimed at improving student well-being.

To evaluate the system's performance, participants will complete mental health assessment questionnaires before and after using the platform, focusing on anxiety, depression, and stress. The accuracy, sensitivity, and specificity of the RL model in identifying emotional states will be assessed, along with changes in the participants' mental health scores. The study's overall effectiveness will be evaluated through a series of experiments and surveys, measuring changes in self-reported mental health and academic performance before and after the intervention. The results are expected to improve mental health support for engineering students by offering personalized, real-time interventions tailored to individual needs.

The bar chart in figure 1 summarizes the count of different emotions across all participants, providing an overview of the most to least frequently recorded emotions. Anger: Most frequently observed with 24 instances. Fear: Notably present, accounted for 17 instances. Happiness: Identified 14 times. Surprise and Disgust: Both emotions were less frequently detected, with 6 and 7 instances each. Sadness: Least frequent, with only 3 occurrences. These insights are based on the emotion label mapping which categorizes the emotions as Anger (0), Disgust (1), Fear (2), Happy (3), Sad (4), and Surprise (5).

Figure 2 Heatmap depicting the correlation between Emotion Day Mean and Overall Mean for various physiological markers (e.g., actigraphy counts, EDA, MET). Warmer colors indicate positive correlations, while cooler colors reflect negative correlations.

3.3.1. Key Findings

The analysis of the relationship between emotional states and physiological markers was conducted by comparing the differences between the emotion day mean and the overall mean across various biometric measures. The results are visualized in the heatmap (Fig.2.), which illustrates the correlations between emotional states (Anger, Fear, Happiness, Disgust, Surprise, and Sadness) and physiological variables, including actigraphy counts, electrodermal activity (EDA), respiratory rate, and other measures.

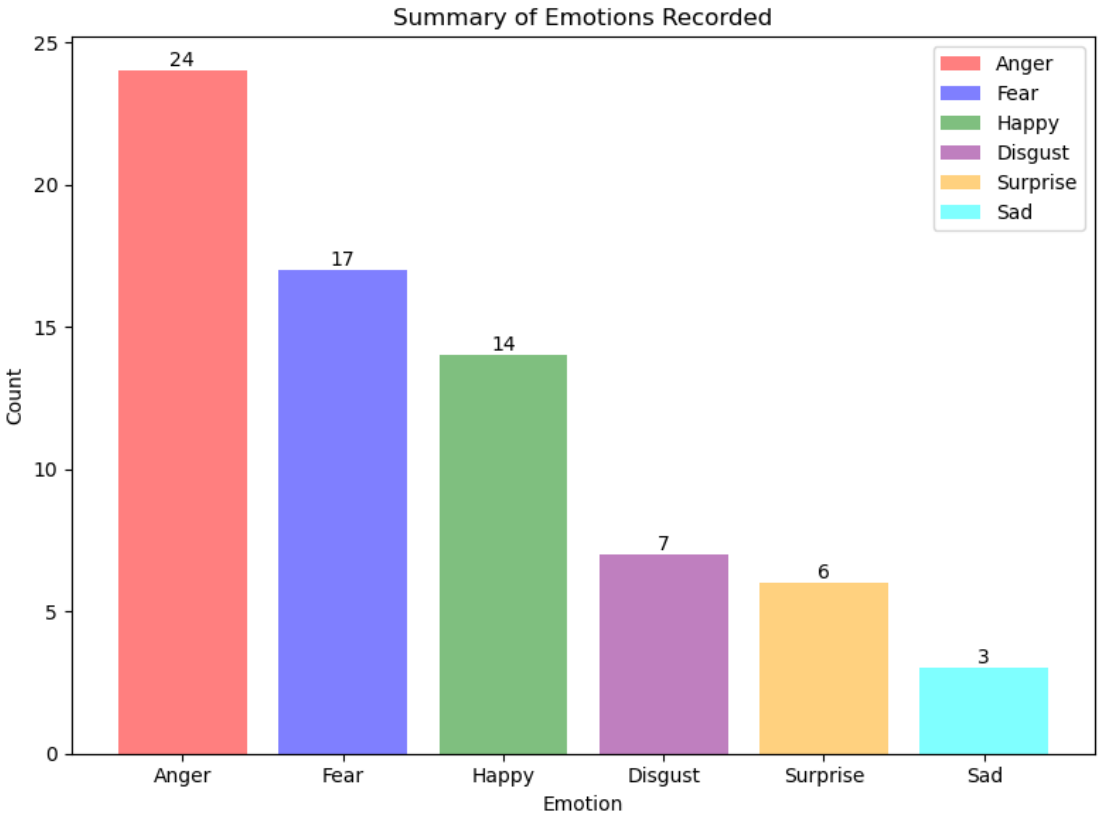


Figure 1: Summary of Emotions Recorded

1. **Electrodermal Activity (EDA)**: A strong positive correlation was observed between EDA and Anger (r = 0.30, p = 0.013), suggesting increased EDA levels during emotional episodes characterized by anger. Conversely, a significant negative correlation was found between EDA and Fear (r = -0.26, p = 0.031), indicating lower EDA levels during fearful states.

2. **Metabolic Equivalent (MET)**: Sadness showed a strong positive correlation with MET (r = 0.32, p = 0.008), suggesting higher energy expenditure on days when participants reported feeling sad.

3. **Activity Counts**: Activity counts demonstrated moderate positive correlations with Anger (r = 0.24, p = 0.046) and Surprise (r = 0.24, p = 0.051), implying increased physical activity during moments of these emotions.

4. **Respiratory Rate**: Sadness also correlated positively with respiratory rate (r = 0.30, p = 0.015), indicating that respiratory rates tended to be higher on days marked by sadness.

5. **Actigraphy Counts**: While most correlations involving actigraphy counts were non-significant, moderate correlations were observed for Sadness across different actigraphy measures, particularly on the Z-axis (r = 0.25, p = 0.038) and the vector magnitude (r = 0.24, p = 0.051), suggesting that emotional states may influence overall body movement.

Overall, the heatmap indicates that physiological changes are indeed associated with distinct emotional states. Anger and Sadness, in particular, exhibit stronger relationships with biometric markers such as EDA, MET, and respiratory rate. These findings demonstrate the potential of using physiological signals to track and identify emotional states in real-time, thus informing targeted mental health interventions for students.

3.4. Limitations of ML and Mental Health

While ML and eye-tracking technologies show significant potential in monitoring mental health, there are some notable limitations. One challenge is the lack of clinical validation, which could hinder its readiness for real-world decision-making in clinical settings. The quality and size of the dataset are also critical factors affecting the performance of ML algorithms. Small sample sizes could lead to overfitting, and if models are only tested within a specific dataset, their generalizability may be limited. Furthermore, ML models often rely heavily on specific input features, meaning their predictions may only be accurate under certain conditions. Studies using binary classifiers also tend to oversimplify conditions and overlook their severity. Additionally, imbalanced datasets often lead to models that predict the majority class while missing rare events.

Future research must focus on recruiting larger, high-quality, and diverse datasets to address the challenges of participant retention and engagement. In this study, although initially recruited 50 participants, the complexity of maintaining engagement throughout the study resulted in 18 active participants by the end. While larger sample sizes might enhance the generalizability of findings, human factors, particularly in mental health studies, introduce complexities that make such generalizations difficult. The current findings should be viewed as a starting point, and

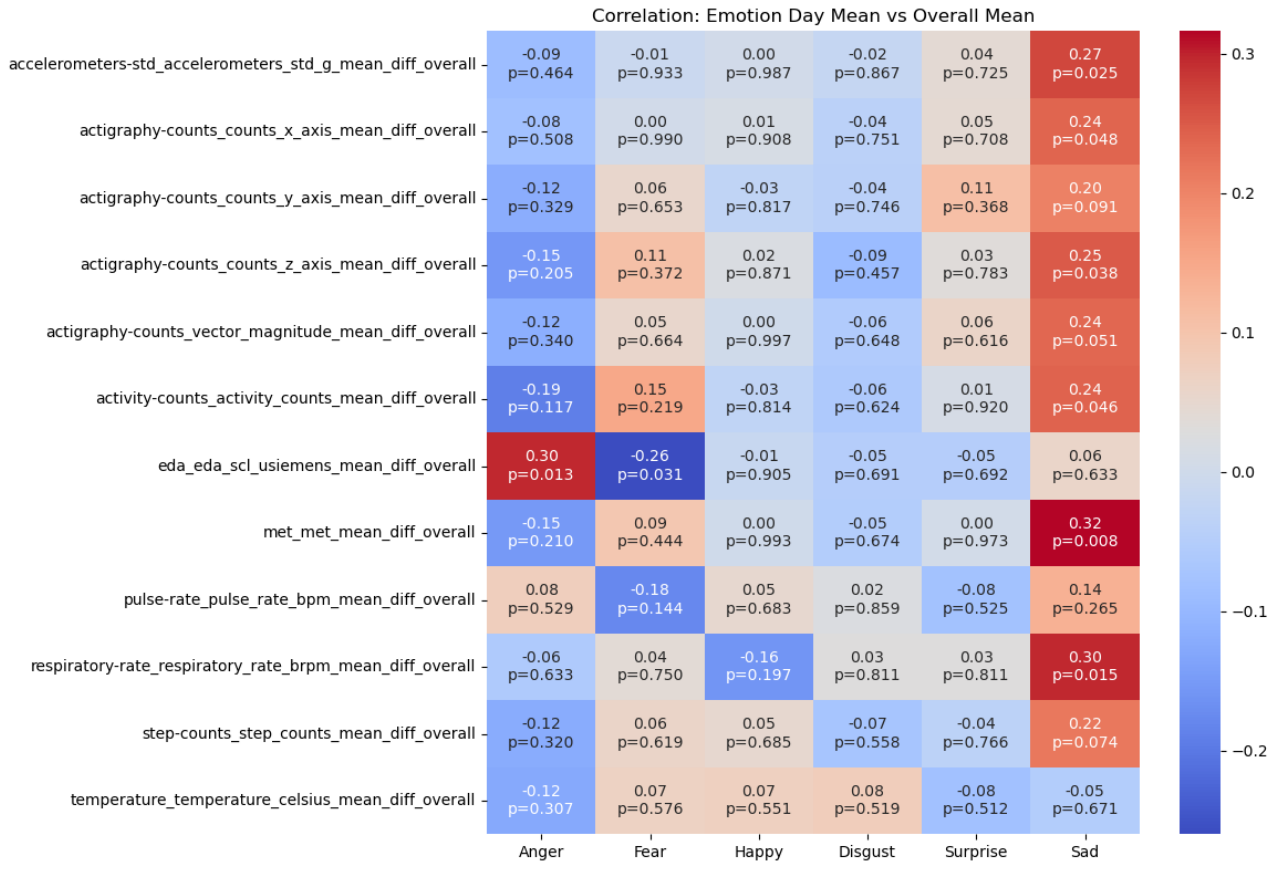


Figure 2: Emotion Day Mean vs. Overall Mean

future research should explore more diverse and larger populations to validate and extend these insights.

Collaboration among institutions for robust data sharing is essential to achieving larger, more representative datasets. Deep learning (DL) methods will become increasingly important in handling complex data, but ensuring that models remain interpretable is crucial. Transfer learning may also be beneficial in improving model performance, especially when working with diverse datasets. Researchers must consider the heterogeneity in input data and develop machine learning models capable of continuous learning to prevent "catastrophic forgetting." Interdisciplinary collaborations between data scientists and clinicians will likely yield the best results for mental health applications, helping to ensure the scalability and reliability of the findings.

3.5. Future Research Directions

Further research should explore a larger sample size and a longer study duration to enhance the generalizability of findings. Larger datasets will allow for more rigorous statistical analysis and reveal insights into the long-term impact of interventions. Personalized feedback and adaptive interventions based on individual stress levels would also improve the system's effectiveness. Expanding the range of physiological measures to include heart rate variability and electrodermal activity could provide a more comprehensive mental health assessment. Finally, researchers should investigate the scalability of the system for broader educational environments, identifying any practical challenges that may arise in larger-scale implementations.

Machine learning is increasingly becoming a cornerstone of digital medicine, with promising applications for mental health. However, to fully unlock its potential, collaboration across disciplines is critical. Clinicians, scientists, and data experts must work together to ensure that ML models are valid, reliable, and free from bias. Moreover, ethical considerations must be addressed, especially when deploying ML technologies in mental health care.

3.6. Recommendations

In conclusion, ML and eye-tracking technologies hold great promise in enhancing mental health support in engineering education. By delivering real-time feedback and personalized interventions, these technologies can help students better manage stress, anxiety, and other emotional challenges, leading to improved academic outcomes and overall well-being. However, further research is necessary to broaden the application of these techniques to other fields and student populations. Ethical concerns, particularly around the use of sensitive personal data in ML applications, must also be addressed. By fostering collaboration between engineers, data scientists, and mental health professionals, the potential benefits of ML in improving student mental health can be fully realized.

Conflict of Interest The authors declare no conflict of interest.

References

[1] R. Baltà-Salvador, N. Olmedo-Torre, M. Peña, A.-I. Renta-Davids, "Academic and emotional effects of online learning during the covid-19 pandemic on engineering students", *Education and information*

technologies, vol. 26, no. 6, pp. 7407–7434, 2021, doi:10.1007/s10639-021-10593-1.

[2] S. Behera, S. S. S. L. Paluri, A. Mishra, "Mental health status of students pursuing professional training: A questionnaire-based study", *Journal of education and health promotion*, vol. 10, no. 1, p. 399, 2021, doi:10.4103/jehp.jehp_1340_20.

[3] J. L. Hopton, S. M. Hunt, C. Shiels, C. Smith, "Measuring psychological well-being. the adapted general well-being index in a primary care setting: a test of validity", *Family Practice*, vol. 12, no. 4, pp. 452–460, 1995, doi:10.1093/fampra/12.4.452.

[4] M. K. Kovich, V. L. Simpson, K. J. Foli, Z. Hass, R. G. Phillips, "Application of the perma model of well-being in undergraduate students", *International journal of community well-being*, vol. 6, no. 1, pp. 1–20, 2023, doi:10.1007/s42413-022-00184-4.

[5] K. J. Hsu, M. E. McNamara, J. Shumake, R. A. Stewart, J. Labrada, A. Alario, G. D. Gonzalez, D. M. Schnyer, C. G. Beevers, "Neurocognitive predictors of self-reported reward responsivity and approach motivation in depression: A data-driven approach", *Depression and anxiety*, vol. 37, no. 7, pp. 682–697, 2020, doi:10.1002/da.23042.

[6] S. Moazemi, S. Vahdati, J. Li, S. Kalkhoff, L. J. Castano, B. Dewitz, R. Bibo, P. Sabouniaghdam, M. S. Tootooni, R. A. Bundschuh, *et al.*, "Artificial intelligence for clinical decision support for monitoring patients in cardiovascular icus: a systematic review", *Frontiers in Medicine*, vol. 10, p. 1109411, 2023, doi:10.3389/fmed.2023.1109411.

[7] B. Mahesh, "Machine learning algorithms -a review", *International Journal of Science and Research (IJSR)*, vol. 9, 2019, doi:10.21275/ART20203995.

[8] T. M. Mitchell, "Machine learning and data mining", *Commun. ACM*, vol. 42, no. 11, pp. 30–36, 1999, doi:10.1145/319382.319388.

[9] R. Sutton, A. Barto, "Reinforcement learning: An introduction", *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054–1054, 1998, doi:10.1109/TNN.1998.712192.

[10] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, "A guide to deep learning in healthcare", *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019, doi:https://doi.org/10.1038/s41591-018-0316-z.

[11] Y. Liu, Z. Zou, A. C. H. Tsang, O. S. Pak, Y.-N. Young, "Mechanical rotation at low reynolds number via reinforcement learning", *Physics of Fluids*, vol. 33, p. 062007, 2021, doi:10.1063/5.0053563.

[12] Z. Zou, Y. Liu, Y.-N. Young, O. S. Pak, A. C. H. Tsang, "Gait switching and targeted navigation of microswimmers via deep reinforcement learning", *Communications Physics*, vol. 5, p. 158, 2022, doi:10.1038/s42005-022-00935-x.

[13] Z. Zou, Y. Liu, A. C. Tsang, Y.-N. Young, O. S. Pak, "Adaptive microlocomotion in a dynamically changing environment via context detection", *Communications in Nonlinear Science and Numerical Simulation*, vol. 128, p. 107666, 2024, doi:https://doi.org/10.1016/j.cnsns.2023.107666.

[14] Y. Liu, B. Zoghi, "Enhancing stem education using machine learning and reinforcement learning techniques for educational software and serious games", pp. 7148–7152, 2023, doi:10.21125/edulearn.2023.1871.

[15] Y. Liu, B. Zoghi, "Emerging technologies in education: Enhancing distance learning with technology-enhanced learning", pp. 7153–7157, 2023, doi:10.21125/edulearn.2023.1872.

[16] Y. Liu, H. Jiang, Z. Ben, "Employing artificial intelligence and machine learning to enhance student learning and outcomes with a focus on building trust and interaction", "EDULEARN24 Proceedings", 16th International Conference on Education and New Learning Technologies, pp. 3069–3074, IATED, 2024, doi:10.21125/edulearn.2024.0814.

[17] Y. Liu, Z. Ben, "Ai-powered strategies for alleviating graduate student burnout through emotional intelligence and wearable technology", "EDULEARN24 Proceedings", 16th International Conference on Education and New Learning Technologies, pp. 3041–3049, IATED, 2024, doi:10.21125/edulearn.2024.0809.

[18] Y. Liu, W. Lu, A. T. Zavareh, M. Rigsby, B. Zoghi, "Improving mental health support in engineering education using machine learning and eye-tracking", "2023 IEEE Frontiers in Education Conference (FIE)", pp. 1–5, IEEE Computer Society, Los Alamitos, CA, USA, 2023, doi:10.1109/FIE58773.2023.10343428.

[19] L. Fierro, Y. Liu, M. Rigsby, B. Zoghi, "Stress and happiness: Investigating stress tolerance and happiness in technical professionals", "INTED2023 Proceedings", 17th International Technology, Education and Development Conference, pp. 3964–3968, IATED, 2023, doi:10.21125/inted.2023.1053.

[20] A. Rehman, S. Abbas, M. Khan, T. M. Ghazal, K. M. Adnan, A. Mosavi, "A secure healthcare 5.0 system based on blockchain technology entangled with federated learning technique", *Computers in Biology and Medicine*, vol. 150, p. 106019, 2022, doi:https://doi.org/10.1016/j.compbiomed.2022.106019.

[21] Z. F. Khan, S. R. Alotaibi, "Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective", *Journal of healthcare engineering*, vol. 2020, no. 1, p. 8894694, 2020, doi:10.1155/2020/8894694.

[22] C. Jyotsna, J. Amudha, "Eye gaze as an indicator for stress level analysis in students", "2018 International conference on advances in computing, communications and informatics (ICACCI)", pp. 1588–1593, IEEE, 2018, doi:10.1109/ICACCI.2018.8554715.

[23] C. Jyotsna, J. Amudha, A. Ram, G. Nollo, "Inteleye: An intelligent tool for the detection of stressful state based on eye gaze data while watching video", *Procedia Computer Science*, vol. 218, pp. 1270–1279, 2023, doi:https://doi.org/10.1016/j.procs.2023.01.105.

[24] A. Kacem, Z. Hammal, M. Daoudi, J. Cohn, "Detecting depression severity by interpretable representations of motion dynamics", "2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)", pp. 739–745, IEEE, 2018, doi:10.1109/FG.2018.00116.

[25] S. Chattopadhyay, "A neuro-fuzzy approach for the diagnosis of depression", *Applied computing and informatics*, vol. 13, no. 1, pp. 10–18, 2017, doi:https://doi.org/10.1016/j.aci.2014.01.001.

[26] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, S. Weidt, *et al.*, "Mobile sensing and support for people with depression: a pilot trial in the wild", *JMIR mHealth and uHealth*, vol. 4, no. 3, p. e5960, 2016, doi:10.2196/mhealth.5960.

[27] L. Avila-Carrasco, D. L. Díaz-Avila, A. Reyes-López, J. Monarrez-Espino, I. Garza-Veloz, P. Velasco-Elizondo, S. Vázquez-Reyes, A. Mauricio-González, J. A. Solís-Galván, M. L. Martinez-Fierro, "Anxiety, depression, and academic stress among medical students during the covid-19 pandemic", *Frontiers in Psychology*, vol. 13, p. 1066673, 2023, doi:10.3389/fpsyg.2022.1066673.

[28] S. P. Behere, R. Yadav, P. B. Behere, "A comparative study of stress among students of medicine, engineering, and nursing", *Indian journal of psychological medicine*, vol. 33, no. 2, pp. 145–148, 2011, doi:10.4103/0253-7176.92064.

[29] S. Pourmohammadi, A. Maleki, "Stress detection using ecg and emg signals: A comprehensive study", *Computer methods and programs in biomedicine*, vol. 193, p. 105482, 2020, doi:10.1016/j.cmpb.2020.105482.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Yuexin Liu has completed her PhD degree in Mathematical Sciences from New Jersey Institute of Technology.

She is currently an instructional assistant professor at Department of Engineering Technology and Industrial Distribution, Texas A&M University, focusing her research on AI and well-being, RL and locomotory design.



Amir Zavareh has completed his PhD degree in Electrical and Electronics Engineering from Texas A&M University.

His research interests are Analog and mixed-signal bioelectronic design, Signal processing, Biomedical imaging and Biomedical instruments.



Ben Zoghi has completed his PhD degree in Biomedical and Electrical Engineering from Texas A&M University.

His research interests are RFID/Sensor technology and how to utilize wearable technology and artificial intelligence (AI)/ machine learning (ML) to develop technical leaders. During his 37 years tenure with Texas A&M College of Engineering, Ben secured over $15M in grants, 9 books and book-chapters and has over 125 peer-reviewed conference and journal papers.

Received: 13 September 2024, Revised: 17 October 2024, Accepted: 18 October 2024, Online: 04 November 2024
DOI: https://dx.doi.org/10.55708/js0311002

Secure Anonymous Acknowledgments in a Delay-Tolerant Network

Edoardo Biagioni*
Edoardo Biagioni, University of Hawai'i at Mānoa Department of Information and Computer Sciences, Honolulu, HI 96822, USA
*Corresponding author: Edoardo Biagioni, 1680 East-West Road, Honolulu HI 96822, USA, +1-808-956-3891 esb@hawaii.edu

ABSTRACT: TCP and many other protocols use acknowledgments to provide reliable transmission of data over unreliable media. **Secure** acknowledgments offer a cryptographic guarantee that valid acknowledgments for a given message can only be issued by the intended receiver. In the context of an ad-hoc network, **anonymous** acknowledgments make it hard for an attacker to determine which device issued a particular acknowledgment. And unlike TCP, the acknowledgments described here work well even for connectionless communications. This acknowledgment mechanism assumes that message data is protected by secure encryption. The sender of a data message includes in the encrypted part of the message a randomly-generated acknowledgment. Only the intended receiver can decrypt the message and issue the acknowledgment. The acknowledgment is issued by sending it out to its peers, who will forward it until it reaches the sender of the data being acknowledged. Such randomly-generated acknowledgments in no way identify senders and receivers, providing a degree of anonymity. This paper describes the use of such acknowledgments in both ad-hoc networks and Delay-Tolerant Networks. In such networks every peer participates in forwarding data, including both the routing and the end-host functionalities of more conventional networks. In a Delay-Tolerant Network, peers may cache messages and deliver them to other peers at a later time, supporting end-to-end delivery even when peers are only connected intermittently. Caches have limited size, so peers must selectively remove cached messages when the cache is full. As an additional aid to selecting messages to be removed from a cache, peers can remove messages for which they have received a matching ack. This can be done while preserving both security and anonymity, by including in every message, unencrypted, a message ID computed as the hash of the message ack sent encrypted with the message. A peer seeing a new ack can then hash it and discard any cached message whose message ID matches the hash of the ack.

KEYWORDS: Ad-Hoc Networks, Delay-Tolerant Networks, Security, Anonymity, Confidentiality.

1. Introduction

The acknowledgement, commonly abbreviated **ack**, is common in networking protocols that seek to provide reliable transmissions over an unreliable medium. TCP relies on acks to confirm receipt of data transmitted on a connection, and also to grant permission to send further data on that connection. Because of this, acks are an essential part of TCP data transmission. TCP acks each have a 32-bit ack number indicating the sequence number of the next byte of data expected on the connection. In a connection transmitting data in just one direction, acks flow in the direction *opposite* the flow of data.

The expected value of a TCP ack can be computed by anyone observing the flow of TCP traffic, so attackers may create and transmit spoofed acks [1].

This paper describes an ack mechanism that provides:

- a guarantee that a message has been delivered, since only the intended receiver can transmit a valid ack for the message (Sections 3 and 4). The intended receiver is securely identified by its ability to decrypt the message.
- a measure of anonymity, in that the ack message in no way identifies either the sending or the receiving device (Section 3). This offers more protection against traffic analysis compared to protocols where the ack carries source and destination addresses.
- the ability to reliably and securely delete acked messages from message caches, even on peers that have no access to any secret key (Section 3).
- communication to the sender that partial receipt of a message is sufficient for the receiver (Section 5).

2. Background

2.1. Secure Hashing

A hash function is a function that computes a fixed-length bitstring, called the hash, from a variable-length input bitstring. In most applications it is beneficial if the hash bears no resemblance to the input. Such hashes are used, for example, in hash tables to evenly distribute the indices to which keys are assigned.

A secure hash or cryptographic hash [2] is a hash that is hard to invert, which means that, given a hash, it is hard to create an input that hashes to that given hash value. Such hashes have many security applications, and several hash algorithms have been standardized by government

agencies such as the US National Institute of Standards and Technology (NIST) [3].

In brief, the result of hashing a bitstring with a cryptographic hash is the hash value or *digest*. As long as the security of the hash is strong, it would be very hard for an attacker who has no access to the original message, to create a bitstring that hashes to the same hash value. In contrast, anyone with access to the original bitstring can easily hash it and verify that it indeed matches the hash value that was received.

The algorithm described in this paper hashes an acknowledgment and sends in the clear the corresponding hash value, which in this paper is called a *message ID*. In the same message as the message ID, the acknowledgment (ack) is sent encrypted, such that only the intended recipient can decrypt it.

After the receiver has decrypted an ack, it can broadcast it. Any peer receiving this decrypted ack can hash it and verify that the previously seen message ID matches the acknowledgment. As long as the hash is hard to invert, only the intended receiver (and the original sender, who also has access to the unencrypted version of the ack) can broadcast a valid ack.

This is an algorithm that can be used with any encryption algorithm and any secure hash function, and in that sense is very general.

2.2. Delay Tolerant Networks

In an ad-hoc network, every peer communicates wirelessly with every other peer in its range. The network is ad-hoc because its connectivity may change with changing conditions. Each peer contributes what it can to data forwarding as well as to creating and receiving data, behaving as a combination of the roles of host and router in other networks.

Delay-Tolerant Network (DTN) technologies [4, 5, 6] support communication among peers that may only occasionally be in communication range of each other. Such conditions are common among mobile peers that communicate through an ad-hoc network. The purpose of DTNs is to deliver data even in the absence of any simultaneous end-to-end path between sender and receiver.

In the 1970s and 1980s email was often delivered even to peers that were never directly connected to the Internet [7]. For example, a peer H (Host) would from time to time dial up another peer G (Gateway) with better connectivity and which had agreed to cache email to and from H. The connection used a protocol called UUCP (unix-to-unix copy) [8] to download emails addressed to the users of H, and to upload emails originated by users of H. H might then in turn forward emails to other peers that depend on it for connectivity.

To support this intermittent email delivery, G had to save messages addressed to users of H and other peers that connected to H, delivering them on request. G would likewise save outgoing messages originating from H and others until G itself could connect to its upstream peer (or directly to the wider Internet) to deliver these messages.

These techniques allowed email to be delivered even if neither sender nor receiver were ever directly connected to the Internet, as long as they were able to connect to someone else with either an Internet connection, or closer to another host that was connected to the Internet.

DTNs have similar goals as the old uucp email system, but delivering general-purpose messages rather than just email. Intermediate peers in a DTN cache messages and deliver them on request. Typically the intermittent connection is established when one of the peers moves into wireless range of another peer.

To accomplish this delivery, in a DTN all devices are peers, so each device must include all the functionality that in a more conventional network is divided among data sources, data sinks, and routers.

When a sending and a receiving DTN node both have access to the Internet, messages can be delivered directly from the sender to the receiver. When the sender, the receiver, or both are not connected to the Internet, they may still be able to communicate directly with each other over ad-hoc links. If at any given time there is no path between sender and receivers, all peers reachable by the sender cache the message, in case they are able to deliver it later.

As connectivity changes and allows communication with new nodes, peers can forward their cached messages to any new peers, with the goal of eventually delivering each message to its final destination.

In their most extreme form DTNs are only useful for data that is not delay-sensitive. Delay-sensitive communications can still occur over ad-hoc networks as long as the devices have an end-to-end simultaneous path over any combination of the Internet and ad-hoc networks, whereas delay-tolerant communications can be supported even without simultaneous end-to-end connectivity.

2.3. A Useful Delay Tolerant Network: AllNet

One useful application of DTNs is delivery of text messages (chat) among mobile devices. Users carry their mobile devices even to locations with no or intermittent connectivity[1]. It would be desirable for users to be able to communicate with at least their neighbors even in the absence of Internet connectivity. Where such connectivity is intermittent, cached messages can be forwarded once connectivity is available. This way of sending text messages resembles at a high level the uucp email delivery described above, but the details are very different, especially the unpredictable availability of wireless channels compared to the scheduled uucp connections over telephone modems.

The benefit of a chat application is that data requirements are moderate, while the usefulness can be very substantial even in situations where delivery of individual chat messages is delayed.

Creating and using ad-hoc networks and DTNs for such applications is the main goal of the AllNet project [9]. AllNet is designed to work whenever devices can communicate directly among each other even in the absence of cellular or Internet service. Available technologies include point-to-point (infrastructure-less) WiFi and the many variants of Bluetooth, particularly Bluetooth Low Energy (BLE). All

[1]The lack of connectivity may be due wilderness adventure, foreign travel, emergency situations, rural areas, or any other reason, such as being outside of the provider's coverage.

of these are available on popular mobile devices, though in many cases the operating system imposes idiosyncratic restrictions on their usage.

Since peers that might forward a message also have the ability to inspect the message, AllNet encrypts the contents of interpersonal messages to prevent eavesdropping. Communication in AllNet needs to identify neither sender nor receiver. Messages may carry optional addresses, only used to improve efficiency of message delivery.

AllNet authenticates users to each other when they are within direct communication range of each other, or when they both know a secret string that allows them to authenticate to each other over the Internet. Such an exchange creates keys that the parties can use at any time thereafter, with assurance that they really are communicating with each other.

AllNet provides some anonymity of communication both with the addresses being optional, and to a lesser extent by the use of ad-hoc communication. AllNet provides such anonymity without being vulnerable to DDoS amplification attacks [10].

The optional nature of addresses in AllNet supports reasonable tradeoffs between anonymity and performance. A message with 0 significant bits of address is anonymous, and so may be delivered to all peers within reach. On the other hand, intermediate peers that care about performance, including especially bandwidth and battery life, may be more willing to forward messages that carry more significant bits of destination address than messages with fewer bits of destination address, since such messages may be delivered more precisely, ultimately consuming fewer network resources. The sender of a message can then choose fewer bits of address to give greater anonymity, or more bits of address to give greater likelihood of message delivery. This choice may be made dynamically, based for example on network traffic, assuming that less overall traffic implies more chances of delivery for anonymous messages.

Ethical Statement: Like every other security feature, encryption, authentication, and anonymity are intended to protect some people from other people. Like every other security feature, anonymity can be used to shield ethical behavior or unethical behavior. This paper does not attempt to distinguish such uses. In general, purely technical work cannot favor ethical over unethical uses of technology.

As is true for many other protocols (including `https`), encryption, anonymity, and authentication are likely to encourage people to use the technology, and their absence would likely discourage people from using technologies such as AllNet. For an ad-hoc network these security properties are essential since there is no assurance that ad-hoc peers, who might forward all of one's messages, will be friendly.

3. Acknowledgments and Caching in an Ad-Hoc Network

Since DTNs work best with all-to-all delivery of anonymous messages, conceptually each peer in a DTN has to cache all messages. Not only do caches have storage limitations, exchanging cached data with every peer that one encounters may require substantial spectrum and too much energy from a limited battery. For all these reasons a peer that caches messages should be informed when one of its cached messages has been delivered to its final destination. AllNet does this by sending a small ack message that confirms receipt for each data message that has reached its final destination.

There are many differences between TCP acks and AllNet acks:

- an AllNet ack gives evidence that an application, rather than the transport layer, has received the message. In TCP, even data that has been acked by the receiving system may never be delivered if the application crashes or stops reading the socket.
- attackers cannot spoof AllNet acks, since only the intended receiver can decrypt the message and issue the corresponding ack. Technically, the sender of the message could also issue the ack, but a legitimate sender by definition is not an attacker.
- in TCP, acks are normal TCP segments that may or may not carry user data. In AllNet acks are 16-byte (128-bit) random strings. As long as the acks are randomly generated, by the birthday paradox the chances of a collision are small as long as there are substantially fewer than $2^{64} = 18,446,744,073,709,551,616$ acks.
- AllNet acks are anonymous. Nothing in the 16-byte random string identifies either the sender or the receiver.
- since an AllNet ack message may carry multiple ack values, a single ack message can acknowledge data messages from different conversations at once. A TCP cumulative ack may acknowledge multiple segments at once, but all such data segments were sent on the same connection.

Each AllNet peer receiving an ack message caches the acks it contains. The peer also hashes the ack, giving the message ID of the message it is acking. If the message ID matches any message that this peer has originated, that message is marked as acknowledged. Also, any matching data message in the peer's cache no longer needs to be cached.

AllNet peers forward ack messages to their ad-hoc peers and across the Internet. Unlike data messages, acks from different messages and from the ack cache may be combined and forwarded together in a single ack message.

When an AllNet peer retransmits a data message for which it has not received an acknowledgment, it may get a matching ack from a peer other than the final receiver, if that peer has cached an ack for that message.

Ack transmission in AllNet is no more efficient as the delivery of the data message, but acks are much smaller than most data messages, so any overhead is less, and likewise the need to evict acks from caches is less – a same-sized cache can hold many more acks than data messages.

Since acks in most systems (including both TCP and AllNet) are idempotent, meaning that receiving the same ack once has the same effect as receiving it multiple times, duplicate transmission of acks has no consequences beyond the cost of transmission.

Acks are also less important than data messages. In contrast to a dropped data message, the worst possible result of

a dropped ack is the sender not knowing that the message has reached its destination and transmitting a duplicate copy of the message, whereas a dropped data message may have the potentially much more serious result of failure to communicate.

4. Technical Details: Secure Acknowledgments

An attacker can easily generate spoofed TCP acks by observing any part of the connection traffic [1]. As a result TCP is not secure, and higher layers such as TLS must be used to provide some assurance of delivery to the intended party.

AllNet provides by design many of the features provided by the combination of TCP and TLS, but for connectionless decentralized ad-hoc networks and DTNs. The security of these transmissions can only be guaranteed if the intended receiver of a message is the only system able issue the corresponding ack.

In AllNet, each receiver holds one or more cryptographic private keys that it uses to decrypt messages addressed to itself. As mentioned above, AllNet certifies keys based on interpersonal interactions among users, whereas TLS relies on hierarchically issued certificates that are vulnerable to hackers penetrating the systems that issue certificates [11].

A sender generating a data message for a specific receiver includes the ack in the plaintext part of the message before encrypting it, at which point the ack is included in the encrypted part of the message.

The sender then adds to the unencrypted part of the data message the cryptographic hash of the ack.[2] This hash is known as the message identifier or **message ID**. The message ID, like the ack, is very likely to be unique for each message.

The entire process is shown in Figure 1.

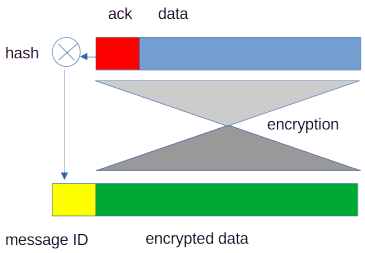


Figure 1: Relationship between ack and message ID

Every peer can see the message ID, but cannot generate a valid ack without access to the receiver's key, so only the intended receiver [3] can generate a valid ack.

A peer that receives an ack can verify whether it matches any of its cached messages or any incoming message by hashing the ack to give the corresponding message ID, then comparing this message ID computed from the ack to the message ID which is in the unencrypted part of every cached or received message.

Since the original sender also has access to the cleartext ack, the sender could also cancel messages that are cached, but this feature is unused both currently and in the foreseeable future.

4.1. Anonymity of Messages and Acknowledgments

Each source and destination address in AllNet messages is defined by a number of significant bits specified as part of the message header. A message whose addresses have no significant bits carries no information about the sender or intended receiver, and as such every peer on the network will try to decrypt it, so that any peer able to decrypt a message is an intended recipient.

The concern with such broadcast messages is the resource usage in having all peers forward and attempt to decrypt every message. The significant bits mechanism of AllNet addresses allows any number of bits to be specified, allowing a sender to specify a small number of bits to reduces the resource usage of the network while still preserving some anonymity. Senders are incentivized to provide as many bits of address as will still retain anonymity, and thereby minimize network resource usage, since packets with more bits of address are more likely to be delivered.

Just as messages can be anonymous, so acks in AllNet are also anonymous in that the ack itself carries no information about the sender and receiver of the ack. In addition, since any peer in the network might have cached the corresponding message, to the extent possible acks are distributed to every peer in the network. This universal distribution makes acks more resistant to traffic analysis than if they were only delivered back to the originator of the data message.

5. Acknowledgments for Messages Larger than the Maximum Transmission Unit (MTU)

The secure acks described so far allow for interesting possibilities when the size of a message is greater than a network's Maximum Transmission Unit (MTU), requiring that the message be sent as a collection of packets instead of a single message. Such large messages can be used to send multimedia data such as audio, images, and video.

The Internet Protocol (IP) uses a mechanism called **fragmentation** [13], and this paper uses the same term. The receiver can reconstruct the larger message once it has received all of the fragments.

In AllNet, each fragment of a larger message carries two encrypted acks, one for the message as a whole and the other for the specific fragment. Correspondingly, the unencrypted part of each fragment contains both a message ID obtained from hashing the message ack, and a fragment ID (which AllNet calls a packet ID) obtained from hashing the fragment ack.

A receiver receiving fragments of a larger message may ack them individually. Once the receiver has received all the fragments of a message, it then issues the ack for the entire message. Each peer receiving a message ack can clear from its cache every fragment it has of the larger message, even if it is only caching some of the fragments.

[2] Normally such hashes produce more than 16 bytes of data, so the sender only includes the first 16 bytes of the hash.

[3] Or anyone who can invert the hash function, which is expected to be challenging for strong cryptographic hash functions such as the SHA-512 hash [12] used by AllNet.

5.1. Acknowledging Partial Transmission

The combination of message acks and fragment acks provides some functionality beyond what traditional TCP segment acks provide. Specifically, if the content can be delivered without delivering all the fragments, then the receiver can issue the message ack immediately even with some fragments still missing.

As one example, email is often sent as both a plain text version and an html version. If either part of the message is received in its entirety, that part can be displayed to the user without having to receive every fragment of the other part, and the receiver can issue the message ack immediately.

When sending image or video data, the resolution of the image or video usually does not need to exceed the resolution of the device. Sending a low-resolution image first, followed by the high-resolution image, allows a low-resolution device to immediately send the message ack, without having to wait for or process the high-resolution image, and likewise allows the sender to only send the low-resolution version. The sender can immediately start sending the high-resolution version if it gets fragment acks for all the fragments of the low-resolution version, but no message ack.

For transmission of more general data (beyond videos and images) there are many schemes for forward error correction (FEC) that involve sending redundant data. If this data can be fragmented appropriately so that the message can be reconstructed even while some fragments are still missing, then a receiver that sends the message ack as soon as it is able to reconstruct the message can avoid having the sender engage in retransmission of any missing fragments.

There are many FEC algorithms. A particularly simple (and inefficient!) FEC scheme simply transmits each fragment 3 times. A receiver that obtains at least one copy of each fragment can immediately issue the message ack, providing reliable transmission without retransmission even in the case of substantial packet loss or transmission delay. Avoiding retransmission is particularly useful in Delay-Tolerant Networks.

If synchronous communication is available, the sender may receive the message ack for such a triply-redundant transmission before sending all three copies of all of the fragments, and can immediately stop transmitting the duplicate/triplicate information. On the other hand in a DTN, a receiver may over time receive a random subset of the fragments, and can then deliver the message to its application and issue the ack as soon as it receives all the fragments needed to reconstruct a complete message. This message ack lets other peers remove from their caches even fragments that the destination has never received.

6. Comparison to TCP Acknowledgments

Section 4 described how AllNet secure acks reliably assure the sender that messages have indeed been received by the intended recipient [4]. This is substantially different from TCP acks, which can be issued by any attacker that knows the sequence numbers in use on the connection and can spoof source IP addresses. A TCP sender has no way to know that such acks are not from the legitimate intended receiver.

Assurance that the ack was issued by the intended receiver is especially valuable in ad-hoc networks and DTNs: since the network is not organized by an authority, there is no reason to believe that intermediate peers are benign. AllNet, as other secure ad-hoc networks, had to be designed assuming that attackers may control at least some of the peers that are forwarding messages and acks.

Further differences between AllNet secure acks and TCP acks follow:

TCP acks carry the sequence number following the last byte that was received. This makes TCP acks cumulative, meaning that a single ack can acknowledge many segment's worth of data, and that loss of a single ack in a continuing stream of data transmission is not likely to lead to retransmission. On the other hand, since TCP acks count bytes rather than messages/packets/segments, and since TCP ack numbers are 32 bits, an ack in the original TCP can acknowledge up to 2^{32} different bytes, or an ack using the Protection Against Wrapped Sequences [14] (PAWS) mechanism can theoretically acknowledge up to 2^{64} different bytes of data.

Unlike TCP acks which count bytes, AllNet secure acks identify packets, so a single 1,000-byte message requires only one distinct secure ack in AllNet but consumes 1,000 sequence and ack numbers in TCP. This is of interest when we consider how many outstanding unacknowledged messages can be handled by each protocol.

The AllNet secure acks are not counters, so cannot be used as cumulative acks (where one ack potentially acknowledges many, many data packets) as in TCP, but one single AllNet message ack can acknowledge many different fragments.

Since there are 2^{128} possible different randomly selected AllNet message acks, the birthday paradox tells us that the chances of collision is extremely low until the number of acks begins to approach $\sqrt{2^{128}} = 2^{64}$.

Unlike TCP, these secure acks might collide with acks from any sender, whereas the TCP connection mechanism ensures that sequence and ack numbers can only overlap within a connection. These differences are summarized in Table 1.

Table 1: Comparison of what TCP and AllNet acks can distinguish.

Protocol	Can reliably distinguish
Original TCP	2^{31} bytes in a window
TCP with PAWS	2^{63} bytes in a window
AllNet	2^{64} simultaneous messages

6.1. Performance Analysis

Given a 64-bit sequence number space for TCP and with reasonable assumptions against delivery of old packets, TCP is guaranteed not to have sequence or ack collisions as long as 2^{63} or fewer bytes are transmitted on a connection within a two-minute Maximum Segment Lifetime (MSL) period, leading to a maximum bandwidth of over 2^{63} / $120s$ = 7×10^{16} bytes/second for each TCP connection.

For the secure acks described in this paper, to stay well away from the birthday paradox we assume that it would be desirable to have no more that about 2^{60} unacknowledged

[4] As long as the recipient's key and the encryption algorithm have not been compromised.

messages in the network at any given time. We further assume message sizes of 1,000 bytes and a maximum message lifetime (as specified by the message expiration option in AllNet) of about a week or 604,800 seconds. Satisfying these assumptions limits the entire network to about $2^{60} \times 1000$ / 604800 = 10^{15} bytes per second.

For communication across the Internet a message lifetime of a week is excessive. Using the same 2 minutes as for TCP, the network can support almost $2^{60} \times 1000$ / 120 = 10^{19} bytes per second.

While this throughput for the entire network cannot be directly compared to the TCP per-connection throughput, it is quite adequate for both the current, largely experimental AllNet, and for any foreseeable developments. Ad-hoc networks are typically small and relatively inefficient [15], and even in the imaginable future are unlikely to scale to large sizes and large amounts of traffic. Therefore for the ad-hoc side of the AllNet communications, even the lower network throughput derived by assuming a 1-week message lifetime is very abundant.

Should the capacity of AllNet ever become an issue due to the limited number of bits in an ack, AllNet could evolve, as TCP already has, to use more acknowledgments bits.

6.2. Performance Results

As in TCP, AllNet acks are sent unreliably, and can therefore be lost. Again as in TCP, the mechanism for requesting an ack retransmission is to retransmit the original message.

Unlike in TCP, for messages that are retransmitted, any intermediate host that has cached the ack can resend it immediately, without the original receiver having to respond. As compared to networks that support TCP and therefore require continuous end-to-end connectivity, for DTNs this increases reliability of ack delivery.

Decrypting a message to generate the ack is more time consuming than the process to generate a TCP ack. TCP has the additional performance advantage of being implemented in the kernel.

With these caveats, we compared the time to return an ack in AllNet between two hosts connected at distant locations across the Internet, with a ping time of 83ms. The first ack (really, SYN+ACK) from the TCP connection establishment 3-way handshake took 125ms (measured using `tcpdump`). Sending a message to an existing contact with AllNet between the same two hosts returned the ack in 154ms (measured using AllNet's `trace` program). These are comparable results.

7. Future Work and Conclusions

It is clear from the above analysis that if AllNet ever becomes as popular as TCP, it will have to be redesigned or extended for higher performance, just as TCP has been and likely will be again in the future. Moving from 16-byte acks to 32-byte acks would address any foreseeable performance limitation due to ack size.

The secure acks described in this paper provide many advantages over conventional acks such as used in TCP. This paper has explored a few, including the guarantee that the ack can only be issued by a receiver that has the correct cryptographic key, the ability to use the combination of message ack and fragment ack to let the sender know how much of the data the receiver actually needs, and the use of the acks to securely enable removing acknowledged messages from peer caches.

Secure acks as described above only work when sent encrypted. Encryption of data messages is pervasive in AllNet, so sending an encrypted ack with the encrypted data of a message requires no additional effort.

Should there be a reason to send the data unencrypted, the ack itself could still be encrypted, as long as keys are distributed in such a way that only the intended recipent can issue the ack.

When encryption is not an option, one could instead imagine having identical ack generators (for example, some kind of secure random number generator) based on identical secret seeds on each pair of sender and receiver, such that the receiver can generate the same sequence of acks as the sender. In such a case, the acks do not need to be transmitted at all. Instead, the sender would include a counter or an identifier for the ack associated with a message, and the receiver can independently generate the ack and hash it to compare the locally generated acks to any received message IDs. The receiver can then issue a valid ack that can be verified by any peer that is caching messages. While these acks do acknowledge receipt, they cannot be used by peers to discard cached messages.

Alternatively, in a scheme somewhat resembling the Bitcoin blockchain [16], and if anonymity is not a concern, a receiver could digitally sign a received message ID with a widely known public key. This scheme requires an infrastructure (perhaps a blockchain?) to record and distribute the public keys, but does not require encryption. Since a shared blockchain requires a persistent connection to the Internet, this scheme is more suitable for systems that can rely on Internet connections than for systems that use ad-hoc and delay-tolerant communications.

We have explored some of the design space for secure and anonymous acknowledgments. While this section has indulged in speculation, the mechanisms in Sections 3 and 4, and in the initial part of Section 5, are fully implemented and live on the AllNet network.

References

[1] Hastings, McLean, "TCP/IP spoofing fundamentals", 1996 International Phoenix Conference on Computers and Communications, doi: 10.1109/PCCC.1996.493637

[2] NIST, "Secure Hash Standard (SHS)", FIPS 180-4, August 2015.

[3] NIST, "SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions", FIPS 202, August 2015. https://csrc.nist.gov/pubs/fips/202/final

[4] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets", SigCOMM, Aug 2003.

[5] Benhamida, Bouabdellah, Challal, "Using delay tolerant network for the Internet of Things: Opportunities and challenges", 2017 8th International Conference on Information and Communication Systems (ICICS), 2017, doi: 10.1109/IACS.2017.7921980

[6] Mallorqui, Zaballos, Serra, "A Delay Tolerant Network for Antarctica", IEEE Communications Magazine, August 2022, doi: 10.1109/MCOM.007.2200147

[7] Partridge, "The Technical Development of Internet Email", IEEE Annals of the History of Computing, vol. 30, no. 2, April-June 2008, doi: 10.1109/MAHC.2008.32

[8] Nowitz, "Uucp Implementation Description", Unix Manual Version 7. https://web.archive.org/web/20180221100921/http://a.papnet.eu/UNIX/v7/files//doc/36_uucpimp.pdf

[9] Biagioni, "Ubiquitous Interpersonal Communication over Ad-Hoc Networks and the Internet", 47th Hawaii International Conference on Systems Sciences), in January 2014, and other papers at https://alnt.org/

[10] Biagioni, "Preventing UDP Flooding Amplification Attacks with Weak Authentication", International Conference on Computing, Networking and Communications (ICNC 2019), February 2019, Honolulu, Hawaii.

[11] Google Security Blog, "An update on attempted man-in-the-middle attacks", August 2011. https://security.googleblog.com/2011/08/update-on-attempted-man-in-middle.html

[12] Penny Pritzker, "Specifications for the Secure Hash Standard", U.S. Federal Information Processing Standards Publication 180-3, October 2008.

[13] Information Sciences Institute, "Internet Protocol", RFC 791 (section 3.2.1.4), September 1981.

[14] Borman, Braden, Jacobson, Scheffenegger, "TCP Extensions for High Performance", RFC 7323, September 2014.

[15] Gupta, Kumar, "The Capacity of Wireless Networks", IEEE Transactions on Information Theory, vol. 46, March 2000.

[16] Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", made public May 24 2009. http://bitcoin.org/bitcoin.pdf

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Edoardo Biagioni did his bachelor's degrees at MIT in 1982, diplom at ETH Zürich in 1985, and PhD degree in computer science at the University of North Carolina, Chapel Hill, in 1992.

He is currently an associate professor in Information and Computer Sciences at the University of Hawai'i at Mānoa, where his research includes networking and systems. He has been working on the AllNet project (www.alnt.org) since 2011.

Received: 11 September, 2024, Revised: 14 October, 2024, Accepted: 15 October, 2024, Online: 07 November, 2024

DOI: https://doi.org/10.55708/js03110003

Fuzzy-Based Approach for Classifying Road Traffic Conditions: A Case Study on the Padua-Venice Motorway

Gizem Erdinc*, Chiara Colombaroni, Gaetano Fusco

Department of Civil, Constructional and Environmental Engineering, Sapienza University of Rome, Rome, 00184, Italy

*Corresponding author: Gizem Erdinc, Dept. of Civil, Constructional and Environmental Engineering Sapienza University of Rome, Italy, +905434493938 & gizemerdincc@gmail.com

ABSTRACT: This study offers a fuzzy-based method for determining the variety of traffic conditions on roads. The fuzzy approach appears more appropriate than the deterministic technique for giving drivers qualitative information about the present traffic condition, as drivers have a shaky understanding of the traffic status. It was used in an analysis that included flow, occupancy, and speed measurements from the Italian freeway that runs between Padua and Venice. MATLAB is used in the application's development. The empirical findings demonstrate how effectively the suggested study performs in classification. The experiment can offer a straightforward and distinctive viewpoint for induction and traffic control on motorways.

KEYWORDS: K Fuzzy traffic congestion estimation, Qualitative traffic state classification, Fuzzy approach

1. Introduction

Transportation plays a crucial role in enabling us to move from one place to another, facilitating the exchange of goods and services, and connecting communities worldwide. There are many vital reasons for the importance of transportation, such as mobility, economic growth, globalization, supply chains, emergency services, tourism, recreation, urban development, and innovation and technology. As a result of the growing economy, people become more inclined to have a car, which leads to jam conditions in traffic on expressways, which is a fundamental element of the transportation system. According to the global economy, traffic congestion may cost billions of dollars by 2023, and this amount tends to grow [1]. Hence, estimating traffic conditions has become a significant problem in transportation engineering.

Severe congestion conditions negatively impact the level of serviceability (LoS) of road networks, leading to poor traffic flow efficiency, environmental degradation, economic loss, and road safety concerns [2]. In this regard, understanding the road traffic conditions as a whole and classifying them may guide traffic engineers to develop effective solutions to corresponding congestion mitigation measures or provide other feasible routes for travelers. Simultaneously, accurate expressway congestion identification can help improve the service level of the road network and encourage the intelligent and refined management of expressways. Therefore, further in-depth research on this topic is required. With in-depth research on traffic conditions, scholars have generally believed that the traffic state has fuzzy characteristics [3, 4]. Defining the traffic flow parameters that mirror the traffic state according to different data characteristics is crucial for traffic state identification [5- 7].

However, it remains a vague notion for which there is currently no agreed-upon definition or measurement. Defined it as a traffic flow state characterized by high density and low speed in [8]. Alternatively, it can be defined as a time-based measure, such as the additional time a vehicle spends in the network when it is unable to navigate at the free-flow speed level in [10] or as a definition of delay in [9]. Conversely, congestion has been interpreted as inadequate road space as an instance of a circumstance where demand surpasses supply on a roadway [11], or when cars collide due to an imbalanced speed-flow relationship [12]. Not to mention, the terms demand-capacity, delay-time, and cost have been used to characterize congestion in over ten distinct ways [13].

Current studies on traffic state identification can be classified into two groups: data- and model-driven. Data-

driven methods can be used for state identification problem without assumptions yet performing only with a large amount of reliable data. Some of common methods include the Bayesian Information Criterion (BIC) [14], Neural Network (NN) [15–16], Support Vector Machine (SVM) [17–19], and K-nearest neighbor (KNN) [20, 21]. While they can predict congestion levels by using artificial intelligence, they are less effective than fuzzy logic for classifying traffic congestion because of its limitations in dealing with real-time uncertainty, adaptability, and data requirements. For example, SVM operates on binary classification (e.g., congested vs. non-congested) or multi-class classification. It assumes clear boundaries between classes, which are not always present in traffic data. This makes SVM less suited for dealing with ambiguous or overlapping congestion levels. On the other hand, adapting a neural network to new traffic conditions requires retraining the model, which can be time-consuming and computationally expensive. Also, their internal workings, especially in deep neural networks, are difficult to interpret, which makes it harder for engineers to understand why the model classifies congestion in a particular way. This lack of transparency can be a major drawback in safety-critical applications like traffic management.

On the other hand, model-driven methods are used to form traffic flow depending on the traffic flow theory, such as Macroscopic Fundamental Diagram (MFD) [22], Lighthill–Whitham–Richards (LWR) [23-27], and Kalman Filter (KF) [28–31] models. Parameters such as speed or density are used to determine the condition of the traffic. Model-driven algorithms exhibit high explanatory power [32]. However, these models generally require idealistic presumptions that might not apply to certain real-world traffic situations and inefficient under fuzzy conditions.

Moreover, traffic is in a state of traffic flow characterized by fundamental variables together, that should be considered in the assessment. In the congested road conditions, the speed may change rapidly, and the congestion estimation based on only the detected speed value may be unstable. Since all three flow variables are known, this application represents the ability of fuzzy logic to handle the relationship between flow, density, and speed. Through all feasible solutions of those variables, the driver's behaviour can be modelled with linguistic definition. For that reason, this paper suggests that the fuzzy methodology might be more appropriate to provide qualitative information to drivers than a simple deterministic threshold-based method because the influence factors of traffic conditions are complicated, and each traffic condition has a certain similarity.

The present paper is structured as follows. After this Introduction, we gave traffic flow characteristics in the Second Section. The Third Section introduces the methodology for quantification. The case study and the findings are discussed afterward. The Conclusion and Discussions section resumes the main assumptions of the methodology implemented and the outcomes produced by the method while comparing them with LOSs by HCM.

2. Traffic Flow Characteristics

Mathematically, traffic is modelled as a flow through a fixed point on the route, analogous to fluid dynamics. In traffic theory, three parameters are typically used to describe traffic flow characteristics: flow (f = vehicle/h), density (k = vehicle/km), and speed (v = km/h).

Flow (f) refers to the number of cars that transit a specific cross-section in a given time unit in a particular interval.

As a measure of the efficiency of highway operations and a direct indicator of how easily cars can move through the traffic stream, density (k) is the characteristic most closely associated with the phenomena of congestion. This symbolizes the quantity of automobiles on a unitary long-road segment at a given time instant. In this study, the density is derived from the loop detectors' occupancy, and the flow and speed values are obtained from the sensors.

Another popular congestion indicator that illustrates the volume of vehicle movement on the road is Speed (v). It should be noted that congestion is a function of speed. Hence, considering speed as an index of congestion is straightforward and intuitive [33].

Equation 1 usually describes the general relationship between density and speed as a linear relationship, as in the Greenshields model [34]:

$$\bar{v}_s = v_l - \frac{v_l}{k_i} k \quad (1)$$

where v_s indicates the space mean speed, v_l indicates the free-flow speed, k indicates the density, and k_i indicates the jam density. The appropriate relationships for the flow and speed can be established as follows using this assumed linear relationship:

$$\bar{v}_s^2 = v_l \bar{v}_s - \frac{v_l}{k_i} f \quad (2)$$

where f is the flow rate. Similarly, the corresponding relationships for flow and density are as follows:

$$f = v_l k - \frac{v_l}{k_i} k^2 \quad (3)$$

Eqs. (2) and (3) show that parabolic relations are found between the flow and density as well as between flow and speed, assuming that the relationship is linear in the shape of Eq. (1) for speed and density. In short, Equation provides a formal expression for the relationship between all of the parameters at the point of equilibrium:

$$f = k * v \quad (4)$$

Traveler information systems find it difficult to give people precise advice since road congestion is not well defined, but a traffic status is characterized with a particular pair of values of the fundamental variables. There isn't a single, unified regulation as of yet [33]. However, there is a vast amount of data on traffic observations, and the phenomenon of traffic is widely recognized. The sole purpose of the example that follows is to serve as an instructive explanation of the ideas behind the fuzzy method that is described in the paper. As is widely known, direct traffic data show two distinct trends in the flow-density plane (Figure. 1a): The flow displays a very noisy and dense structure throughout the range of higher values of density, with an average declining tendency as the density increases. In the range of low levels of density, the flow is practically linear with density and presents few deviations from the average speed. It is commonly recognized that the high-density regime is typified by erratic flow conditions that are dictated by the microscopic mechanism that underlies the traffic stream. These fluctuations include even minor driving movements that result in a stop-and-go traffic pattern.

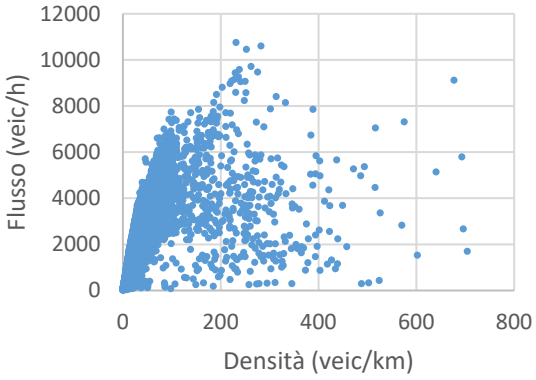


Figure 1a: Density-Flow relationship [35]

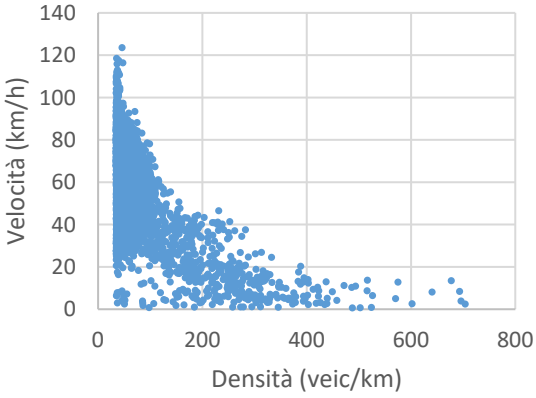


Figure 1:. Density-Speed relationship[35]

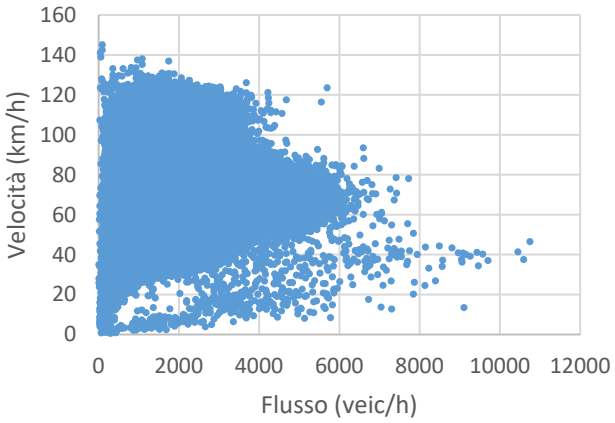


Figure 1c: Flow-Speed relationship [35]

Figures 1a, 1b and 1c illustrate a 15-min observation of the traffic of one of the sections for several months and it is evident from the figures that collecting traffic data for a long period continuously introduces a huge noisy component. The scope of the proposed model is to simulate the general state of the road and derive a relationship between the fundamental variables by applying the fuzzy Mamdani inference approach regardless of their value [33]. Each part has been modelled and executed independently because they could each have a distinct general state. The network consists of two main branches: the North (Highway) and the South (Tangential), the results of which are provided in Section 4.

3. Methodology

A traffic stream can be characterized by its current state and the change in its condition. To identify in which state the traffic is in, congestion should be quantified and detected how it is spreading. From this point of view, this paper is an evaluation of the average situation of traffic process in the related time interval as an index between [0-1]. The main purpose is to answer the questions *what is happening*, and *how critical is the current situation in terms of congestion condition level* which reflects the severity of traffic. It has been built with a fuzzy model in MATLAB.

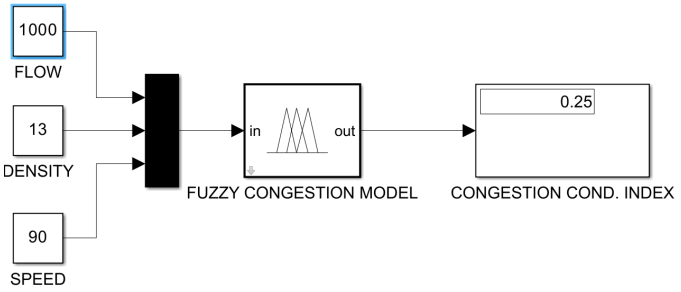


Figure 2: Evaluation of the current traffic condition.

The fuzzy model named 'Fuzzy Congestion Model', represented in the middle of the Figure 2, computes the average congestion level by using average flow, average density, and average speed information. All input and output variables have been defined by setting linguistic

terms which help us represent them with a wide range. We set $n_{x1} = n_{x2} = n_{x3} = 5$ linguistic terms as:

$$F = \{FF, RFF, AF, CF, VCF\},$$
$$K = \{VLD, LD, MD, HD, VHD\},$$
$$V = \{VS, S, A, F, VF\}.$$

The output variable is also classified as:

$$CCI = \{VL, L, M, H, VH\}$$

Explanation of variables are: FF: Free Flow, RFF: Reasonably Free Flow, AF: Average Flow, CF: Congested Flow and VCF: Very Congested Flow, VLD: Very Low Density, LD: Low Density, MD: Medium Density, HD: High Density and VHD: Very High Density, VS: Very Slow, S: Slow, A: Average, F: Fast and VF: Very Fast, VL: Very Low, L: Low, M: Medium, H: High and VH: Very High.

While linguistic definitions give the variables a thinking and understanding way to close humans, they need to represent numerically. In order to complete the process, all variables are converted from crisp numbers into membership degrees of the fuzzy set, or fuzzified. Membership functions are utilized in this process, providing the numerical value's degree of belongingness to the associated fuzzy set inside a closed interval [0,1]. Here, 1 expresses full membership and 0 states non-membership. In this paper, triangular membership (Equation 5) functions are used, since they are one of the most used examples and they are effective in reflecting the characteristics of the fuzzy sets used here.

$$\mu(x) = \begin{cases} 0, x \leq \alpha min \; or \; x \geq \alpha max \\ \frac{x-\alpha min}{\beta-\alpha min}, \; x \in (\alpha min, \beta) \\ \frac{\alpha max-x}{\alpha max-\beta}, \; x \in (\beta, \alpha max) \end{cases} \quad (5)$$

As an example of the mathematical form, the membership function of the *Medium level of Density (MD)* variable is provided in Eq. 6 and its demonstration is given with Figure 3.

$$\mu MD(k) = \begin{cases} 0, k \leq 11 \; or \; k \geq 30 \\ \frac{k-11}{20-11}, \; x \in (11,20) \\ \frac{30-k}{30-20}, \; x \in (20, 30) \end{cases} \quad (6)$$

Figure 3 shows the membership function plots of the density input variable. All the Very Low, Low, Medium, High, and Very High-density levels are the set members of K. Hence, the *MD* is a set member of the K fuzzy set. It has a membership degree of K fuzzy set that Eq. 6 gives us. This means it has a certain quantity of numerical belongingness to the K fuzzy set. That numerical value in [0-1] is calculated by Eq. 6.

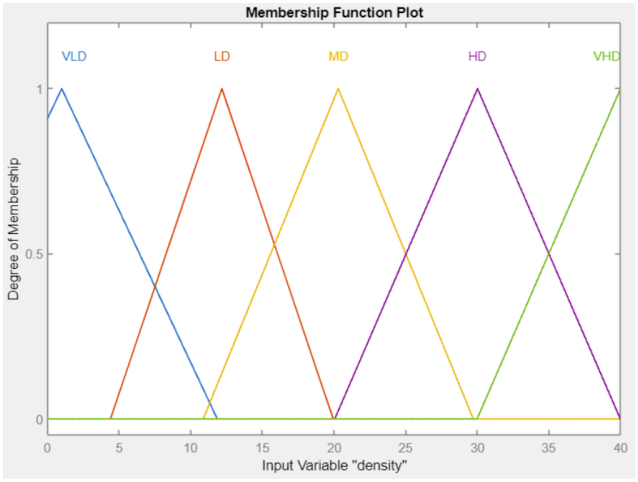


Figure 3: Membership functions for density variable

Similarly, Figure 4 shows the membership function plots of congestion condition level index output variable. In both plots (Figure 3 and 4), the horizontal plane shows the value range of the related variables, while the vertical plane gives the membership function value of it in the range [0-1]. The range intervals of the related variables are set after a close investigation according to the data. It is to note that the terms are imprecisely defined following the fuzzy logic concept, and there are no clear cuts between the fuzzy congestion levels.

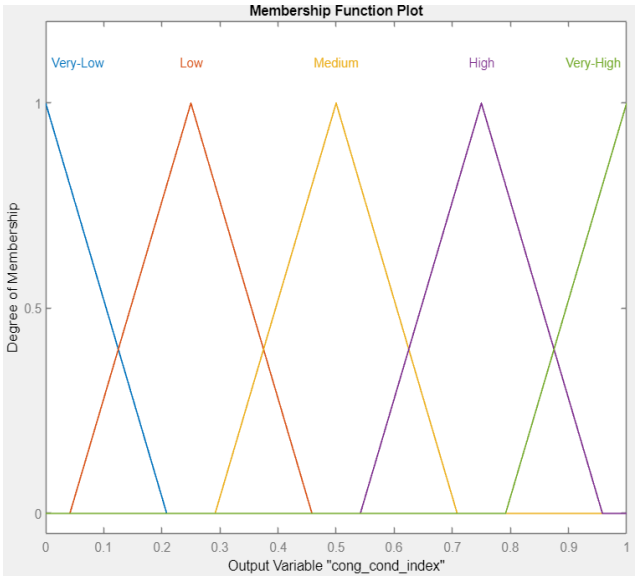


Figure 4: Membership functions for congestion condition variable

The fundamental aspect of the approach is to represent the input-output connection, which shows how input variables reflect onto the output universe, in order to construct the nonlinear surface model and make an inference. We define 45 rules for this application and they can be stated mathematically as:

$$(F \bullet F(f)) \; \theta min \; (K \bullet K(k)) \; \theta min \; (V \bullet V(v)) \; \rightarrow \; (A \bullet A(f, k, v))$$

where the symbol • states the linguistic term is, symbol θmin states the logic operator AND. Some of fuzzy if-then rules are given in Table 1 below. 45 rules have been modelled for this application.

Table 1: If-then rules for congestion condition index

IF	AND	AND	THEN
Flow is Free Flow	Density is Low	Speed is Fast	CCI is Very Low
Flow is Free Flow	Density is Medium	Speed is Average	CCI is Low
Flow is Congested	Density is High	Speed is Slow	CCI is High
Flow is Very Congested Flow	Density is Very High	Speed is Very Slow	CCI is Very High

The framework of the method is given in Figure 5. The figure shows the structure of the model, with input parameters, assigned linguistic variables to our crisp inputs, applied fuzzy rules and operators, chosen implication and aggregation method as well as calculated value of congestion as a defuzzied output.

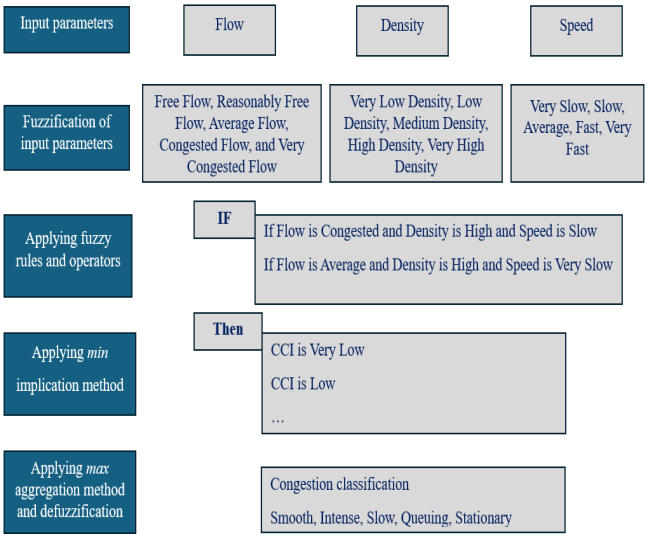


Figure 5: General framework of the study

Detailed information about methodological steps such as variable definitions, fuzzification, rules formulations and defuzzification can be seen in [36].

4. Application Results

The main scope of this study is to identify the traffic condition ranges of Padua-Venice motorway network, which is composed by 4 branches of 3-lane roads with separate carriageways, the motorway has a total length of about 74 km and 16 inter-sections. Real traffic data collected from 31st December 2018 to the 30th of August 2019 contains the following information: local unit code, code section counting, day type, road (section) id, date, flow, density, and harmonic speed aggregated every 15 minutes. The whole data was examined statistically section-by-section; however, it is important to mention that, due to temporary failures of some detectors, some parts of the data are missing. We have neglected them in the application step. In this research, we investigated the congestion from Monday through Saturday for eight months, during the hours of 6:30 am to 9:00 am. The application is run on MATLAB fuzzy logic toolbox R2020b and SIMULINK.

4.1. Traffic situation in the motorway (Northern part)

The first step of the application consisted in computing the speed predictions for 15-minute time intervals between 6:30-9:00 for all sections along the northern part of the network. Table 2 shows the congestion levels with average observed intervals, and Figure 6 depicts the traffic condition for all sections of the motorway.

A similar pattern, with intense or slow state levels, may be seen in the observations of traffic states with average intervals along the entire freeway. At section 43, the change to a slow traffic status is smooth. Sections 42 and 44, on the other hand, move to a more intensive condition. The point at which an intense condition transitions to a slow state is reached when flows exceed 1400 cars per hour and density passes 17 vehicles per kilometres.

Table 2: Traffic state results with average observed intervals

Section number	Road section	Average of the Observed Intervals			Congestion Condition Index	Traffic State
id	id	Flow (vehicle/hour /lane)	Density (vehicle/km /lane)	Observed Speed (km/h)	[0-1]	
24-1	742308	602	7	101	0.21	Intense
3-1	742456	1534	21	88.2	0.52	Slow
15-1	742250	1534	21	88.2	0.52	Slow
4-1	742463	930	11	86.6	0.25	Intense
5-1	742240	848	10	88.9	0.25	Intense
6-1	742483	1775	24	87.5	0.53	Slow
16-1	742476	1775	24	87.5	0.53	Slow
17-1	742230	1775	24	87.5	0.53	Slow
18-1	742215	2368	27	92.5	0.61	Slow
19-1	742496	2368	27	92.5	0.61	Slow
9-1	742202	1578	20	87.4	0.52	Slow
10-1	742508	1578	20	87.4	0.52	Slow
20-1	742512	1768	22	83.3	0.55	Slow
21-1	742194	1768	22	83.3	0.55	Slow
41-1	741467	1531	17	97	0.38	Slow
42-2	740855	1480	15	107	0.27	Intense
42-1	741527	1475	15	108	0.27	Intense
43-1	741555	1453	17	93	0.37	Slow
44-2	740700	1145	14	106	0.25	Intense
44-1	741682	1042	13	105	0.25	Intense
45-1	741708	1432	15	99	0.25	Intense
46-1	741741	973	11	98	0.24	Intense

In this case, the average speed decreases by around 15% from its previous level. This means that drivers can drive at around 67% of the free-flow speed level of the road. In such a traffic state, flow and density rates decrease to around 1150 vehicles/h and 14 vehicles/km, respectively. This allows the speed level to grow up to around 106.0 km/h, which equals more than 75% of its free-flow level, and it gives drivers more freedom to manoeuvre. There is no severe congestion and no Stationary state along the highway with this subset of data, because having an average level of flow and low level of density allows Fast level of speed with a high membership and Average and Very Fast levels of speed

with low memberships, so this situation occurs in an Intense state generally. It was coloured in Figure 6.



Figure 6: The condition of the motorway using a subset of data for the hours of 6:30 am to 9:00 am

4.2. Traffic situation in the highway (Tangential part)

The same computations have been done for all sections in the Tangential part of the network as well. This application is run with the same subset of data with 5-minute time intervals to criticize the effectiveness of the timeframe in characterizing the trend of rate changes. The observations of traffic states with average intervals all along the highway show a similar pattern, with intense or slow state levels. However, after having an intense situation, there is a clear transition to a slow state at sections 3 and 15 with an almost double amount of congestion indices. In this case, the average levels of flow and density go up to 2.5 and 3 times higher than the previous section's levels and that intensity causes a medium congestion and it is classified in the next congested state by the proposed model. Like the motorway part, there is no severe congestion and no Stationary state along the highway, as vehicles can move in a Fast level of Speed with a range between 85.5 km/h to 101km/h with a low membership, so this situation occurs in a slow state generally. Identified traffic states have been colored in Figure 7.

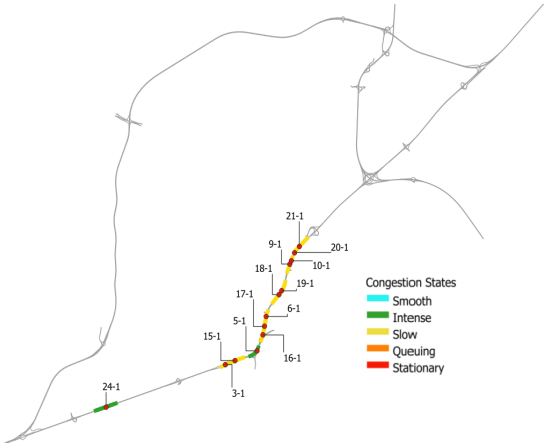


Figure 7: The condition of the highway using a subset of data for the hours of 6:30 am to 9:00 am

5. Discussion and Conclusion

5.1. Evaluation of congestion index by the fuzzy congestion calculator with flow, density, and speed variables

Table 3 summaries the results in evaluating congestion condition index (CCI) for some selected intervals of the field data including variables for density, speed, and flow. The intervals are chosen at random so that the traffic conditions encompass the entire spectrum of the traffic level index. The following lists the LOSs established by the Highway Capacity Manual [37] are shown for comparison. The detailed comparison with the LOSs was given in the next section.

Here, Table 3's observation (with a focus on average values) enables the following important conclusions to be drawn:

- Table 3 demonstrates that there is a comparatively large variation of density for the same speed. For instance, the densities for the sampling numbers 8, 9, 17, 24, and 99 at the speed level with v = 88-89 (km/h) are 8, 20, 18, 42, and 22 (veh. /lane/km), respectively.
- Similarly, there is a comparatively large range in speed for a given value of density. For instance, sample numbers 46, 52, and 89 have a speed range of [118–77] (km/h) for density k = 31 (veh./lane/km), but their flow range (veh./lane/hr) is rather wide (3588, 495, 375). These show that the traffic state cannot be adequately represented by a single variable.
- Additionally, Table 3 demonstrates the speed is not highly responsive to density alone at the "Very Low" and "Low" levels. For densities k = 2, 3, and 12, the corresponding average speeds (veh./lane/km) are 111.3, 106.4, and 99.0 (km/h). Speed drops by only 11.05% as densities rise six times from 2 to 12 (veh./lane/km). While the flow increases along with the density, the speed also reduces dramatically. For instance, when the flow hits 1800–2000 (veh. /lane/hr), as in indices of 69, 73, and 89, speed drops 26.6% to the level of 73–77 (km/h). This corroborates the idea that the traffic status is a specific tuple of the basic variables' values. As such, it requires contemporaneous attention.

The congestion indices generated by the FIS system are displayed in Figure 8 across with the LOS as described by [37]. Despite the fact that LOSs are connected to a comparatively broad range of congestion condition indices (CCIs), Figure 8 illustrates a general correlation between the CCIs and service levels. The CCIs of the various LOSs significantly overlap. It is possible to state that Very Low congestion (μVL = {-0.208 0 0.208}) is mainly linked with LOS A and partially with LOS B in

Table 3: Evaluation results

No	Section id	Unix time	f (veh/ hour /lane)	v (km/h)	k (veh/ km /lane)	CCI	traf_ state	LOS
1	742308	450	1872	112.2	17	0.390	slow	D
2	742202	0	1044	87.9	12	0.250	intense	C
3	740855	1140	2480	114.2	22	0.581	slow	F
4	742250	0	648	113.0	6	0.168	intense	A
5	742215	0	1155	96.1	12	0.250	intense	C
6	742496	0	520	101.5	5	0.157	intense	A
7	742230	400	1368	87.7	16	0.335	slow	D
8	742508	0	708	88.6	8	0.237	intense	B
9	742230	395	1716	88.4	20	0.500	slow	E
10	742250	80	120	100.7	1	0.899	smooth	A
11	742463	5	168	111.3	2	0.080	smooth	A
12	742496	1285	1520	97.0	16	0.375	slow	D
13	742250	0	312	90.7	4	0.100	smooth	A
14	742476	495	3696	75.0	51	0.500	queuing	F
15	742496	520	1440	97.8	15	0.356	slow	D
16	742194	1430	1296	90.6	14	0.306	slow	C
17	742240	945	1572	88.8	18	0.411	slow	D
18	742463	640	732	84.3	9	0.242	intense	B
19	742483	5	468	90.3	5	0.223	intense	A
20	742240	1435	1000	90.2	11	0.250	intense	B
21	742456	0	240	99.0	3	0.092	smooth	A
22	742496	245	600	78.7	8	0.224	intense	B
23	742202	415	2064	89.6	23	0.504	slow	E
24	742483	1050	3708	88.5	42	0.752	stationary	F
25	742508	1260	780	82.2	9	0.246	intense	B
26	742508	0	3084	18.9	25	0.678	queuing	F
27	741741	1065	852	112.0	8	0.215	intense	B
28	742512	0	270	90.4	3	0.100	smooth	A
29	742476	805	1812	76.8	24	0.589	slow	F
30	742194	1430	1692	82.1	21	0.538	slow	F
31	742240	1435	336	93.1	4	0.098	smooth	A
32	742308	0	360	107.3	3	0.083	smooth	A
33	742463	95	324	52.4	9	0.218	intense	B
34	742308	450	2424	110.4	22	0.575	slow	F
35	742230	395	1176	99.0	12	0.254	intense	C
36	742215	0	1440	90.7	16	0.356	slow	D
37	742463	610	1716	82.8	21	0.537	slow	F
38	742308	450	1290	108.7	12	0.297	intense	C
39	742496	520	3600	92.5	39	0.750	queuing	F
40	742215	0	2220	90.9	25	0.530	slow	E
41	742202	0	108	100.7	2	0.089	smooth	A
42	742476	1065	2892	76.4	38	0.774	queuing	F
43	742230	400	768	99.2	8	0.231	intense	B
44	740765	30	316	106.4	3	0.084	smooth	A
45	742215	5	600	95.6	6	0.216	intense	B
46	741527	510	3588	118.1	31	0.750	queuing	F
47	740700	150	64	125.0	4	0.071	smooth	A
48	742512	350	936	69.1	14	0.250	intense	C
49	742508	680	3048	71.8	43	0.785	stationary	F
50	742240	1430	444	93.0	5	0.223	intense	B

relation to the fuzzy partitioning of the CCI in Figures. 8 and 4. Low congestion (μL = {0.041 0.25 0.458}) is mostly linked to LOS C, sporadically to LOS B, and barely to LOS A. A little amount of medium congestion (μM = {0.30 0.5 0.708}) is linked to LOS C, and most of it to LOS D and E. Finally, Very High congestion (μVH = {0.791 1 1.208}) is strictly associated with LOS F.

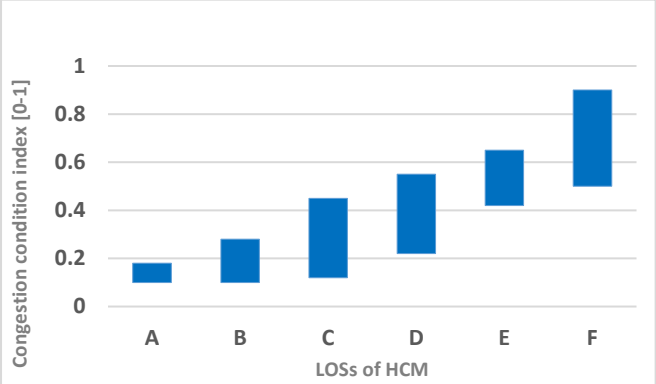


Figure 8: Correspondence between congestion condition levels and levels of service defined by HCM.

5.2. Comparison with the US Highway Capacity Manual's Service Level definitions

To provide additional insight into the significance of the suggested approach, the outcomes have been contrasted with the similar LOSs estimations derived using the HCM method [37], which establishes service traffic levels using the figure 9 diagram as a basis. Since the HCM speed-flow curves derive from observations conducted in the USA, they should only be used as a means of comparing different approaches and applied very carefully to an Italian motorway. The speed limit on the freeway under investigation is 130 km/h, which is higher than the ranges taken into account by the HCM. On the other hand, a LOS can be applied to the actual traffic states if a projection of the patterns of HCM curves is accepted. This experiment, however, shows that the combined observations of all three state variables are inconsistent with the theoretical model, which focuses on a univariate relationship between flow and speed and the state equation (3). It's true that if density is taken into account as the LOS identification indicator, then the recorded state of traffic on the road falling between 11 veh/km/lane (18 veh/mi/lane) and 17 veh/km/lane (27 veh/mi/lane) would be classified as either LOS C or LOS D. Nevertheless, the associated LOS will change from LOS B to LOS C if the flow values—which span an array of values from 973 veh/h/lane to 1531 veh/h/lane—are taken into account. More importantly, the density ranges from 7 veh/km/lane to 27 veh/km/lane on the highway would be associated with all levels from A to F with density guidance, while they would be considered in levels B to F for a huge range of flows between 602 veh/h/lane and 2368 veh/h/lane. These traffic conditions were categorized using the fuzzy technique as either slow or intense. The fuzzy method overcomes classification ambiguity caused by inherent mistakes created by the steady-state principles underpinning the HCM methodology and encompasses the inevitable ambiguity of traffic state classification because it takes into account all three state variables at once.

Furthermore, because the qualitative description makes use of common descriptors and is better suited for informing drivers, it is easier than a scale classification. A sensitivity analysis was conducted in terms of the condition of traffic range index between [0-1] to observe the impact of the chosen membership function of the suggested fuzzy model. For that analysis, five cases with triangular, trapezoidal and Gauss, Gaussbell, and Gauss2 function shapes have been carried out and the triangular membership function is stepped out as the best function type [36].

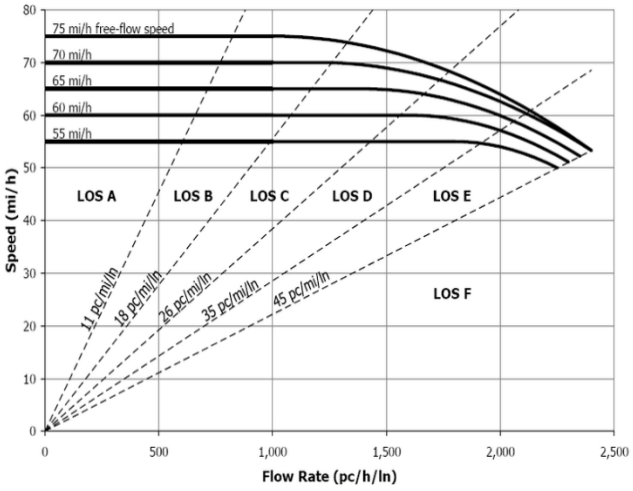


Figure 9: Speed-flow-density relationship [37]

5.3. Conclusion

This article handles the traffic state identification problem and proposes a Mamdani-based fuzzy logic application with a much bigger real dataset from a motorway in Italy. To address the value of multi-dimensional evaluation and we already possess all three-traffic flow variables, we examined the capability of fuzzy reasoning to simulate the relationships between variables and driving behaviour modelling. Moreover, predicted states have been discussed by comparing the LOSs by HCM. The fuzzy technique, which is based on a non-univocal range characterization of traffic circumstances, looks to be a more acceptable strategy for traffic congestion classification because LOS is a stepwise assessment and does not convey a concept of congestion level clearly. Since the fuzzy method considers all three

state variables, it takes into account the inevitable uncertainty in traffic state recognition and removes any ambiguity in categorization caused by inaccuracies resulting from the steady-state assumptions that underlie the HCM methodology. Furthermore, because the qualitative description makes use of common descriptors and is better suited for informing drivers, it is more straightforward than a scale classification. When properly designed using rules, the fuzzy-based congestion analyser can be used as a tool to help control actions on expressways under boisterous conditions. The current applications are in places like Singapore and Los Angeles [38-39]. These systems continuously monitor the state of the roads and modify the length of the signals. As a result, traffic flows more smoothly, stop-and-go situations are less frequent, and fuel consumption is decreased. Moreover, some Indian towns are experimenting with fuzzy logic-based systems that enable traffic signal modifications in real-time upon detection of emergency vehicles [40]. This facilitates the clearance of space for emergency vehicles and, by cutting reaction times in urgent situations, may save lives. By incorporating intelligence into traffic management, these solutions enhance effectiveness, security, and environmental sustainability.

However, due to the case study's static environment, this analysis limited to show severe and/or temporal congestion. For future works, the model will be extended with a consideration of this issues. Additionally, unanticipated events like accidents, weather changes, etc., will be handled to make developing control actions on motorways more convenient.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has not been supported by any foundation.

References

[1] Y. Liu, C. Zhen, D. Huili, "Highway traffic congestion detection and evaluation based on deep learning techniques." *Soft Computing* vol. 27, no.17, 12249-12265, 2023, doi.org/10.1007/s00500-023-08821-6.

[2] C. Wang, C. Yuting, W. Jie, Q. Jinjin, "An improved crowddet algorithm for traffic congestion detection in expressway scenarios." *Applied Sciences*, vol. 13, no. 12, 2023, 7174, doi.org/10.3390/app13127174.

[3] T. G. Huang, L. H. Xu, and X. Y. Kuang. "Urban road traffic state identification based on fuzzy C-mean clustering." *Journal of Chongqing Jiaotong University (Natural Science)*, vol. 34, no. 2, 101-106, 2015.

[4] H.W. Liu, Y.T. Wang, X.K. Wang, Y. Liu, Y. Liu, X.Y. Zhang, F. Xiao, "Cloud Model-Based Fuzzy Inference System for Short-Term Traffic Flow Prediction", *Mathematics*, vol. 11, no. 11, 2509, 2023, doi.org/10.3390/math11112509.

[5] Z. Liu, C. Lyu, Z. Wang, S. Wang, P. Liu, Q. Meng, 2023 "A Gaussian-process-based data-driven traffic flow model and its application in road capacity analysis." *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, 1544-1563, 2023, doi.org/10.1109/TITS.2022.3223982.

[6] K. Zhao, G. Dudu, S. Miao, Z. Chenao, S. Hongbo, "Short-term traffic flow prediction based on VMD and IDBO-LSTM." *IEEE Access* 2023, doi.org/10.1109/ACCESS.2023.3312711.

[7] J, Ying, "Research on methods of urban road traffic condition identification", *Qingdao Univ. (Eng. Technol. Ed.)*, vol. 27, no. 3, 84–87, 2012, doi.org/10.1007/978-981-19-1057-9_26.

[8] P.H. Bovy, I. Salomon, "Congestion in Europe: Measurements, patterns and policies." *Travel behaviour: spatial patterns, congestion and modelling*, 143-179, 2002, doi.org/10.4337/9781781951187.

[9] A. Almeida, S. Brás, S. Sargento, I. Oliveira, "Exploring bus tracking data to characterize urban traffic congestion". *Journal of Urban Mobility*, vol. 4, 100065, 2023, doi.org/10.1016/j.urbmob.2023.100065.

[10] M.A.P. Taylor, "Exploring the nature of urban traffic congestion: concepts, parameters, theories and models." In *Proceedings, 16th Arrb Conference, 9-13 November, Perth, Western Australia*, Vol. 16, No. 5. 1992.

[11] I. Pedram, S. Salek, R. Omrani, *Defining and Measuring Urban Congestion—Guidelines for Defining and Measuring Urban Congestion*, Transportation Association of Canada: Ottawa, ON, Canada. 2017, Access (December 2020).

[12] ECM, *The spread of congestion in Europe*, 11th Round Table on Transport Economics, OECD Publication Service, 1999.

[13] M. Aftabuzzaman, "Measuring traffic congestion - a critical review", *Proceedings of the 30th Australasian Transport Research Forum*, 16, 2007.

[14] J. Xia, W. Huang, J. Guo, "A clustering approach to online freeway traffic state identification using ITS data", *KSCE Journal of Civil Engineering*, vol. 16, 426-432, 2012, doi 10.1007/s12205-012-1233-1.

[15] B. Priambodo, A. Azlin, A.K. Rabiah, "Prediction of average speed based on relationships between neighbouring roads using K-NN and neural network.", 18-33, 2020, doi.org/10.3991/ijoe.v16i01.11671.

[16] N. Tilahun, P. Thakuriah, "Incorporating weather information into real-time speed estimates: Comparison of alternative models", *J. Transp. Eng.*, vol. 139, no.4, 379–389, 2012, doi.org/10.1061/(ASCE)TE.1943-5436.0000506.

[17] L. Zhang, J. Ma, X. Liu, M. Zhang, X. Duan, Z. Wang, "A novel support vector machine model of traffic state identification of urban expressway integrating parallel genetic and C-means clustering algorithm", *Tehnički vjesnik*, vol. 29, no. 3, 731-741, 2022, doi.org/10.17559/TV-20211201014622.

[18] Y. Zhu, Y. Zheng, "Retracted article: traffic identification and traffic analysis based on support vector machine." *Neural Computing and Applications*, vol. 32, no. 7, 1903-1911, 2020, doi.org/10.1007/s00521-019-04493-2.

[19] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification." *Expert Systems with Applications* vol. 176, 114885, 2021, doi.org/10.1016/j.eswa.2021.114885.

[20] D. Mladenović, S. Janković, S. Zdravković, S. Mladenović, A. Uzelac, "Night traffic flow prediction using K-nearest neighbors algorithm", *Operational research in engineering sciences: theory and applications*, vol. 5, no. 1, 152-168, 2022, doi.org/10.31181/oresta240322136m.

[21] F. Harrou, Z. Abdelhafid, S. Ying, "Traffic congestion monitoring using an improved kNN strategy." *Measurement*, vol. 156, 107534, 2020, doi.org/10.1016/j.measurement.2020.107534.

[22] T. Seo, T. Kusakabe, Y. Asakura, "Calibration of fundamental diagram using trajectories of probe vehicles: Basic formulation and heuristic algorithm", *Transportation Research Procedia*, vol. 21, 6-17, 2017, doi.org/10.1016/j.trpro.2017.03.073.

[23] C.G. Claudel, A.M. Bayen, "Guaranteed bounds for traffic flow parameters estimation using mixed Lagrangian-Eulerian sensing," *46th Annual Allerton Conference on Communication, Control, and Computing*, 636–645, 2008, doi.org/10.1109/ALLERTON.2008.4797618.

[24] J. Jiang, N. Dellaert, T. Van Woensel, L. Wu, "Modelling traffic flows and estimating road travel times in transportation network under dynamic disturbances", *Transportation*, vol. 47, 2951-2980, 2020, doi.org/10.1007/s11116-019-09997-3.

[25] G.B. Whitham, *Linear and Nonlinear Waves*, Wiley-Interscience, 1974.

[26] R. Pradhan, S. Shrestha, D. Gurung, "Mathematical modeling of mixed-traffic in urban areas." *Mathematical modeling and computing*, vol. 9, No. 2, 226-240, 2022, doi: 10.23939/mmc2022.02.226.

[27] A. Iftikhar, Z.H. Khan, T.A. Gulliver, K.S. Khattak, M.A. Khan, M. Ali, N, Minallahe, "Macroscopic traffic flow characterization at bottlenecks", *Civil Engineering Journal*, vol. 6, no. 7, 1227-1242, 2020, doi.org/10.28991/cej-2020-03091543.

[28] J. Junfeng, H. Wang, "Traffic behavior recognition from traffic videos under occlusion condition: a Kalman filter approach." *Transportation research record,* vol. 2676, no. 7, 55-65, 2022, doi.org/10.1177/03611981221076426.

[29] F. Zheng, S.E. Jabari, H.X. Liu, D. Lin, "Traffic state estimation using stochastic Lagrangian dynamics", *Transportation Research Part B: Methodological*, vol. 115, 143-165, 2018, doi.org/10.1016/j.trb.2018.07.004.

[30] Y. Liu, Z. Cai, H. Dou, "Highway traffic congestion detection and evaluation based on deep learning techniques", *Soft Comput*, vol. 27, no. 20, 12249–12265, 2023, https://doi.org/10.1007/s00500-023-08821-6.

[31] A. Nantes, D. Ngoduy, A. Bhaskar, M. Miska, E. Chung, "Real-time traffic state estimation in urban corridors from heterogeneous data", *Transportation Research Part C: Emerging Technologies*, vol. 66, 99-118, 2016, doi.org/10.1016/j.trc.2015.07.005.

[32] C. Wang, W. Zhang, C. Wu, H. Hu, H. Ding, W. Zhu, "A traffic state recognition model based on feature map and deep learning" *Physica A: Statistical Mechanics and its Applications*, vol. 607, 128-198, 2022, doi.org/10.1016/j.physa.2022.128198.

[33] T.D. Toan, Y.D. Wong, "Fuzzy logic-based methodology for quantification of traffic congestion", *Physica A: Statistical Mechanics and its Applications*, vol. 570, p.125784, 2021, doi.org/10.1016/j.physa.2021.125784.

[34] A. Skabardonis, P. Varaiya, K.F. Petty, "Measuring recurrent and nonrecurrent traffic congestion", *Transportation Research Record*, vol. 1856, no. 1, 118-124, 2003, doi.org/10.3141/1856-12.

[35] Erdinç, G., Colombaroni, C. and Fusco, G., "Two-stage fuzzy traffic congestion detector", *Future transportation*, vol.3, no. 3, 840-857, 2023, doi.org/10.3390/futuretransp3030047.

[36] G. Erdinç, C. Colombaroni, G. Fusco, "Application of a Mamdani-Based Fuzzy Traffic State Identifier to a Real Case Study", In *2023 8th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* 1-6, IEEE, 2023, doi.org/10.1109/MT-ITS56129.2023.10241428.

[37] National Academies of Sciences, Engineering, and Medicine, *Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis*, Washington, DC: The National Academies Press, 2022, doi.org/10.17226/26432.

[38] G. Hewei, A.S. Sadiq, M.A. Taheir, "Adaptive fuzzy-logic traffic control approach based on volunteer IoT agent mechanism", *SN Computer Science*, vol. 3, no. 1, p.68, 2022, doi.org/10.1007/s42979-021-00956-3.

[39] B. Yulianto, "Adaptive Traffic Signal Control Using Fuzzy Logic Under Mixed Traffic Conditions", *Proceedings of the 5th International Conference on Rehabilitation and Maintenance in Civil Engineering. ICRMCE 2021. Lecture Notes in Civil Engineering*, vol 225. Springer, Singapore, doi.org/10.1007/978-981-16-9348-9_59.

[40] A. Agrahari, M.M. Dhabu, P.S. Deshpande, A. Tiwari, M.A. Baig, A.D. Sawarkar, "Artificial Intelligence-Based Adaptive Traffic Signal Control System: A Comprehensive Review", *Electronics*, vol.13, no. 19, p.3875, 2024, doi.org/10.3390/electronics13193875.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Gizem Erdinc has done her bachelor's and master's degree from Erciyes University in 2014 and 2016 respectively. She completed her first PhD at Erciyes University in 2020 and the second one at Sapienza University of Rome in 2024.

She has experience in fuzzy logic applications in civil engineering problems, transportation planning and construction management.

Chiara Colombaroni has done his bachelor's, master's and PhD degree from Sapienza University of Rome.

Her research activities are addressed to transportation planning and transportation system modelling, driver behaviour modelling, traveller information systems, road traffic control and network design.

Gaetano Fusco has done his degree from Sapienza University of Rome. He is a Full Professor of Transport at the Faculty of Engineering of the University of Roma "La Sapienza.

Received: 18 September, 2024, Revised: 17 October, 2024, Accepted: 18 October, 2024, Online: 08 November, 2024

DOI: https://doi.org/10.55708/js0311005

A Comparative Analysis of Interior Gateway Protocols in Large-Scale Enterprise Topologies

Saleh Hussein Al-Awami *, **Emad Awadh Ben Srity** , **Ali Tahir Abu Raas**

University of Benghazi, Department of Data Communications and Computer Networks, Benghazi, 00218, Libya

*Corresponding author: Saleh AL Awami, Department of Software Engineering Faculty of Information Technology, University of Benghazi ,Benghazi, Libya, & saleh.awami@uob.edu.ly

Author 2 Email: itstd.4074@uob.edu.ly , Author 3 Email: itstd.3522@uob.edu.ly

ABSTRACT: Interior gateway protocols (IGPs) have gained popularity in networking technologies due to their capacity to enable standardized and flexible communication among these algorithms. In autonomous systems (AS), network devices communicate with one another via IGPs. This work presents a fresh investigation into the performance of inner gateway protocols in large-scale enterprise topologies. Also, the experiment lab using GNS3 simulation has been conducted to evaluate and examine the performance of RIP, EIGRP, OSPF, and IS-IS, taking into account convergence time, latency, and jitter in large-scale network topology. The work has used a tri-connected architecture, with ten (10) routers connected via three serial connections and fifteen (15) network subnets, resulting in thirty (30) different paths for routing data packets for each tested routing technique. End-to-end delay, jitter, and convergence time are three measurement measures used to investigate network topologies. The experiment's outcomes have revealed that EIGRP has superior delay and convergence time performance. Furthermore, the results have been showed that IS-IS outperformed OSPF in terms of convergence time. Overall, this work improves the field by giving a grand average computation approach for measuring the jitter metric, which has been compared to a standard method. The method has been thoroughly explored utilizing derivate statistical equations and associated pseudo code.

KEYWORDS: RIP, EIGRP, OSPF, IS-IS, Interior gateway protocol, convergence time, In-depth analysis.

1. Introduction

Interior Gateway Protocols (IGPs) are routing protocols used to exchange routing information within an autonomous system (AS), such as a corporate network or a campus network. IGPs are essential for ensuring efficient and reliable communication among the routers and hosts within an AS. However, as the size and complexity of enterprise networks increase, so do the challenges and requirements for IGPs.

The process of routing involves determining the most efficient path for the flow of traffic within one or more networks. Routers utilize specific protocols that enable them to exchange relevant information on remote networks and update their routing tables regularly. These routing protocols determine the best path for each network, showcasing the interconnections among all routers within the network. This, in turn, facilitates communication primarily among neighboring routers and eventually throughout the entire network [1]. Routing can be classified into two categories: static and dynamic, as shown in Figure 1.

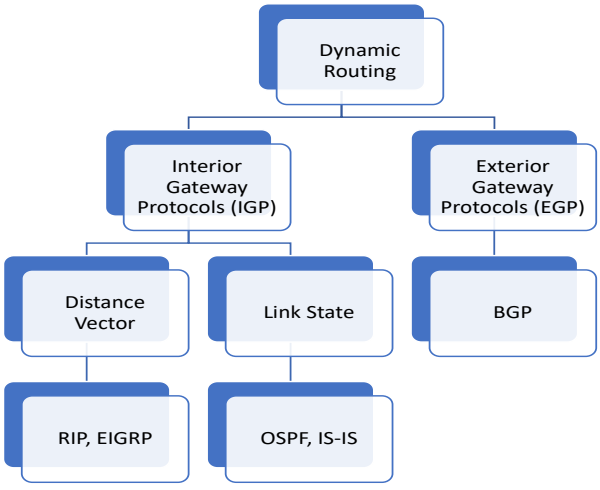


Figure. 1:Classified of Routing Protocols.

Static routing refers to the manual configuration of IP addresses and the manual entry of routes into the routing table by the network administrator before the routing process starts. Such routes can only be modified by the administrator and do not adapt to changes in the network, making them less suitable for large and unpredictable networks. They do not consume CPU memory or link

bandwidth, nor do they transmit information that can be intercepted by hackers:[1].

In contrast, dynamic routing utilizes routing protocols that allow the network administrator to input the protocol syntax instead of manually entering each IP address. These protocols are advantageous and practical in real-time as they can detect changes within the network and determine the shortest path. They read routing update messages and adjust to changing network conditions by recalculating the shortest path and transmitting the message:[2].

This article is organized into several sections. Section 2 covers Interior Gateway Protocols, while Section 3 is about related works. Section 4 discusses the problem statement, and Section 5 presents the proposed work. Section 6 describes the network topology model and measurement methodology. Section 7 presents the experiment results and proposed mathematical equations and comparing to the basic conventional one, and section 8 presents the discussion and Section 9 concludes the overall work of this article.

2. Interior Gateway Protocols

This section provides an overview of dynamic routing protocols in local area networks: Dynamic routing protocols can be classified into three main types: distance-vector, link-state, and hybrid. Distance-vector protocols rely on periodic exchanges of routing information between neighboring routers. Each router advertises the distance and direction to reach a destination network to its neighbors. Distance-vector protocols do not require routers to have a complete knowledge of the entire network topology. These protocols involve sending all or a portion of a router's routing table to adjacent routers. Examples of distance-vector protocols are RIP (version 1 and 2) and IGRP. Link-state protocols, on the other hand, require routers to collect link state information from all routers in the network and construct a network map based on that information. Then, each router calculates the best routes to each network using the network map. OSPF and IS-IS are examples of link-state protocols. Hybrid protocols combine the features of both distance-vector and link-state protocols.

2.1. RIP

Routing Information Protocol (RIP), a distance vector routing protocol, uses hops as the primary criterion for selecting a quality route. A maximum of 15 hops is allowed for RIP. RIP is also known as the Ford-Fulkerson or Bellman-Ford algorithm. There are three main versions of RIP: RIP v1, RIP v2, and RIPng. RIP information is carried in UDP packets. The port number for RIP V1 and V2 is 520 and the port number for RIPng is 521.The administrative distance assigned to the RIP protocol is 120. Routers that use RIP send their routing tables to their neighbors every 30 seconds. The updates are sent to the multicast address 224.0.0.9 for V1 and V2. The updates for RIPng are sent to the multicast address FF02 :: 9.

2.2. EIGRP

The Enhanced Interior Gateway Routing Protocol (EIGRP) is a hybrid routing protocol that combines the features of link-state and distance-vector protocols. It uses various criteria such as delay, bandwidth, reliability, and load to select the optimal path. It employs the Diffusion Update Algorithm (DUAL) for fast convergence and effective route optimization. It also uses the Reliable Transport Protocol (RTP) for delivering EIGRP packets. There are two versions of EIGRP: EIGRPv4 for IPV4 and EIGRPv6 for IPV6. The administrative distance of EIGRP is 90 by default, with a maximum hop count of 255. It sends packets to the multicast address 224.0.0.10 and uses the port number 88. Internal routes have an administrative distance (AD) of 90, while external routes have an AD of 170. The EIGRP metrics are computed using the constants K1 = 1, K2 = 0, K3 = 1, K4 = 0, and K5 = 0, as shown in (1), and in (2):.[3]

$$Metric = \left[K1 \times bandwith + \frac{(K2 \times bandwith)}{(256 - load)} + K3 \times delay\right] \times \left[\frac{K5}{reliability + K5}\right] \times 256 \quad (1)$$

$$\text{Metric} = 256 \times (Bandwith + delay) \quad (2)$$

2.3. OSPF

The Open Shortest Path First (OSPF) routing protocol utilizes Dijkstra's algorithm, a link-state routing algorithm. Dijkstra's algorithm calculates the shortest paths to all destinations from a source node, rather than a single source and destination pair, differing fundamentally from the Bellman-Ford algorithm and distance vector routing protocols. The routing metric in OSPF depends on the cost of a link. OSPF does not restrict the number of hops in a route. OSPF exists in two primary versions for IPv4(OSPFv2) and IPv6 (OSPFv3) networks, operating under the same principles. In OSPF, each router generates Link State Advertisements (LSAs) to establish and maintain a consistent representation of the routing topology across the entire domain. OSPF routers collect the network's link state data and store it in a Link State Database (LSDB). Each OSPF router determines the shortest path to each network segment using the Shortest Path First (SPF) algorithm. The well-known port for OSPF is 89. OSPF uses an administrative distance of 110. OSPFv2 utilizes multicast address 224.0.0.5, while OSPFv3 uses multicast address FF02::5. The OSPF metrics are illustrated in (3) [3].

$$Cost = \frac{10\text{^}8}{\text{bandwith}} \ (bps) \quad (3)$$

2.4. OSPF

The Intermediate System to Intermediate System (IS-IS) routing protocol operates similarly to OSPF. IS-IS is a link-state protocol developed by the International

Organization for Standardization (ISO) [4]. Like OSPF, IS-IS utilizes a link state database and Dijkstra's Shortest Path First (SPF) algorithm to determine the shortest path routes. IS-IS supports four distinct routing metrics: default, delay, expense, and error.

IS-IS does not limit the number of hops in routes and has an administrative distance of 115. IS-IS routers can be classified as level 1, level 2, or level 1-2 routers. Level 1 routers do not have direct connections to other areas, while level 2 routers connect multiple areas, similar to OSPF Area Border Routers (ABRs). Level 1-2 routers function as both level 1 and level 2 routers, connecting their area to the backbone.

3. Related Works

This study [2] focuses on the comparison of routing protocols used in computer networks and their performance metrics, including end-to-end delay, throughput, and convergence duration. The study utilizes the Optimized Network Engineering Tool (OPNET MODELER) to simulate different network topologies and compare the performance of three routing protocols: RIP, EIGRP, and OSPF. The simulation results show that OSPF has the fastest throughput among the three protocols, while EIGRP has the fastest convergence time. Additionally, EIGRP has faster throughput than RIP, but RIP has the highest queuing delay. This study provides valuable insights into the performance of different routing protocols and their suitability for different network topologies, but the methods of calculating performance metrics were not clarified, and the Riverbed Modeler they relied upon.

This study [5] compares the performance of EIGRP and OSPF routing protocols in terms of convergence time with link failures and the addition of new links in different network scenarios. The research uses a network simulator called GNS3 to simulate network topologies and validates the results using Cisco hardware equipment in the laboratory. The study finds that EIGRP has a faster convergence time than OSPF with a link failure or a new link added to the network. The experiment contributes to existing knowledge by identifying that the mesh topology has the best convergence time and that hardware implementations of routing protocols are better than using a network simulator. The study recommends further research in comparing BGP with EIGRP and OSPF and analyzing the protocol changes or adaptations in terms of convergence time with the versions of IPV4 and IPV6. Additionally, latency and quality of service are identified as vital areas of research in both EIGRP and OSPF routing protocols. The Wireshark results are used to check the network configuration and monitor accurate time responses for the various packets.

This study [6] compares the performance of two routing protocols, RIPv2 and EIGRP, in terms of data packet routing speed and convergence time. The research uses the Cisco Packet Tracer 7.10 software to simulate full mesh and half mesh topologies and conducts various tests, such as network connectivity, traceroute, and time testing. The study finds that EIGRP has a faster convergence time and routing speed than RIPv2 in both topologies. The results of the analysis suggest that EIGRP's hybrid routing characteristics (distance vector and link state) enable faster selection of vector-based paths and that EIGRP updates only the routing table of affected routers, while RIPv2 updates the routing table of all routers. The study concludes that EIGRP is a better routing protocol for an institution or college network that requires proper and safe network methods.

This study [7] aims to assess and compare the performance of different interior gateway routing protocols in an enterprise-level network that is complicated and operates in real-time. This is achieved by utilizing the GNS3 software to simulate the network, which consists of hosts, switches, and routers, and by putting each protocol into the topology. The protocols that have been examined are EIGRP, OSPF, and RIP. The authors guarantee secure data transmission at each node by offering authentication, and the assessment criteria include throughput, end-to-end delay, and convergence time. The study's findings indicate that EIGRP is the most practicable interior gateway routing protocol for a complicated enterprise-level network since it performs the best in terms of convergence time. In stable network conditions, there are no major differences observed in the end-to-end delay and throughput values of the three protocols; OSPF has a higher throughput than RIP and less significant delays than both RIP and EIGRP.

This study [8] aims to assess and contrast the effectiveness of many Interior Gateway Routing Protocols (IGRPs), more especially OSPF, RIPV2, and EIGRP. Using Graphical Network Simulator (GNS-3) simulations, the authors analyzed a number of characteristics, including throughput, jitter, convergence time, end-to-end delay, and packet depletion. EIGRP performs better than OSPF, according to the results. It was also mentioned that EIGRP uses greater processing power, which increases the system's power usage.

This study [6] the convergence times of the EIGRP and RIPv2 routing protocols were compared by the authors. Cisco Packet Tracer 7.10 was used to run the simulation on two different network topologies: full mesh and half mesh. The findings showed that the RIPv2 protocol took longer to converge—ranging from 0.01 to 0.19 seconds—while the EIGRP protocol had a faster average convergence time of 0.01-0.02 seconds.

4. Problem Statement

Network hardware performance is influenced by several factors, including convergence, delay, and jitter. Achieving optimal performance requires selecting the appropriate combination of routing protocols. Despite extensive research on interior gateway protocols, detailed computation of convergence, delay, and jitter has not been sufficiently addressed in these studies. In contrast, our paper employs equations to calculate these factors, providing a more comprehensive and accurate evaluation of network hardware performance, then the proposed analysis has .

5. Proposed Work

The primary aim of this work is to evaluate and analyze the performance of a complex enterprise network by considering convergence, delay, and jitter, and utilizing Interior Gateway Protocols such as OSPF, EIGRP, and IS-IS. Furthermore, the study seeks to determine the most suitable protocol for large-enterprise topologies. Notably, this research contributes to the field by introducing an enhancement analysis of the "delay, jitter, which is a factor in achieving optimal network performance. In order to obtain a higher accuracy measured value than the fundamental traditional method, the grand average statistical method has been utilized in the proposed method to quantify jitter, "delay", in a large-scale network topology. In order to determine the differences in the jitter, "delay", in the proposed Tri-connected network devices topology in this work, a comparison is made between the suggested approach and the conventional one using the interior gateway routing protocols.

6. Models and Measurement Parameters

This section provides an overview of the devices and media types employed in setting up the network topology, as well as an examination of the software used. Additionally, the measurement parameters used to analyze the network topologies are described.

Partial mesh (Tri-connected) topology is a network topology combining features of full mesh and star topologies, providing high fault tolerance and redundancy. It is commonly used in medium to large enterprises, offering a scalable traffic handling capability. Nodes are connected to a subset of other nodes or all nodes in the network. It is ideal for medium to large enterprises requiring a network that can handle a moderate to a high amount of traffic. Partial mesh topology is less expensive than a full mesh topology. It is an effective solution for enterprise networks that require fault tolerance, redundancy, and scalability.

The proposed network topologies comprise ten routers interconnected through three serial connections, with 15 network subnets forming 30 multiple paths for routing data packets for each assessed routing protocol. This configuration results in a complex network topology at the enterprise level, as illustrated in Figure 2 below.

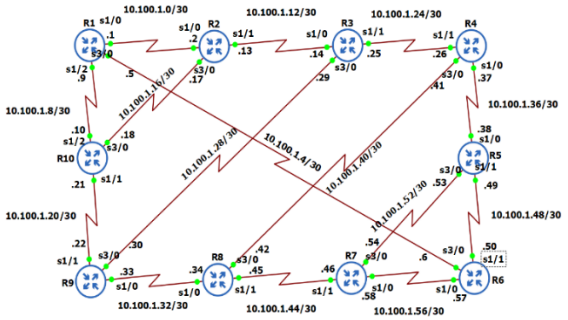


Figure. 2: Large-scale enterprise network topology.

The proposed large-scale infrastructure is built to ensure it ideal for enterprise applications and the complexity of network devices. The network configuration is built to meet the following:

A) **Network topology**: ten (10) routers connected via three (3) serial connections to interconnect using a single channel to routing packets.

B) **Subnets:** fifteen (15) different subnets to ensure the complexity of the topology.

C) **Routing path**: thirty (30) routing path for data packet routing (Tri-connected 10 routers).

D) **Routing protocol**: various interior routing protocols assessed for efficiency.

E) **Network topology Size**: Large-scale network topology

6.1. Hardware and software resources

This subsection provides details on the hardware and software resources utilized for the simulation and testing. The IP addressing configuration and network topology used in the experiments are illustrated in Figure 2 above. The network topology comprises ten Cisco 7200 routers, providing multiple paths between different sections of the topology. The routers used in the experiments are identical in model and specifications, as listed in Table 1. The software used to perform the experiments is GNS3 version 2.2.34, while the HP ProBook 450 5G was used as the testing device, with its specifications outlined in Table 2.

Table 1: Network devices employed in the network topology

Factor	Description
Device Model	Cisco 7200 router
Processor	NPE- 400
RAM Memory	512MB
Flash Memory	32 MB

Table 2: Laptop device specification for Personal Computer use.

Factor	Description
Model	HP ProBook 450 5G
Processor	Intel(R) Core (TM) i5-8250U
RAM Memory	8G
CPU Speed	CPU @ 1.60GHz
Operating System	Windows 11 Pro

From the standpoint of three initial parameters, network performance was examined using the same network topology and different protocols.

The measurement parameters would be used in this work can be defined as follows:

A) End to end delay

The total time required for a complete message to travel from its source to destination encompasses the duration from the initial transmission of the first bit of the message until the final delivery of the last bit of the message.

B) Jitter

The variation in packet delay, commonly referred to as packet delay variance, is an indication of the presence of jitter when data packets experience disparate delays across a network. Jitter is often measured in milliseconds (MS).

C) Convergence time

The metric measures the speed at which a group of routers restores a network to its normal state.

7. Experiment Results

In this section, each interior gateway protocol presented in this paper was evaluated to determine the most suitable protocol for the network topology under consideration. This evaluation was based on the calculation of end-to-end delay, convergence time, and jitter, in accordance with the study's objectives.

7.1. End-to-End Delay Testing Results

Table 3 presents the average end-to-end delay values for EIGRP, OSPF, and IS-IS, which were determined using the ping utility throughout the analysis.

Table 3: The average end-to-end delay values for the RIP, EIGRP, OSPF, and IS-IS protocols calculated for each subnet

Subnet #	IP Address	RIP	EIGRP	OSPF	IS-IS
1	10.100.1.1/30	170	40	48	42
2	10.100.1.5/30	196	42	24	50
3	10.100.1.9/30	152	42	40	48
4	10.100.1.13/30	83	24	26	32
5	10.100.1.17/30	26	20	22	20
6	10.100.1.21/30	21	20	22	25
7	10.100.1.25/30	40	39	32	30
8	10.100.1.29/30	48	42	57	27
9	10.100.1.33/30	35	31	28	42
10	10.100.1.37/30	57	56	66	64
11	10.100.1.41/30	77	68	86	52
12	10.100.1.45/30	46	58	65	60
13	10.100.1.49/30	40	40	38	42
14	10.100.1.53/30	44	38	36	37
15	10.100.1.57/30	26	21	24	19

Table 3 represents end to end delay for each protocol by taking the average of round-trip time (RTT) by using ping utility. Figures. 3, 4, and 5 represent samples of results of sixty (60) experiments for four (4) measured interior gateway routing protocol for fifteen (15) different subnets.



Figure 3: The Ping utility executed on 10 routers utilizing the RIP protocol.

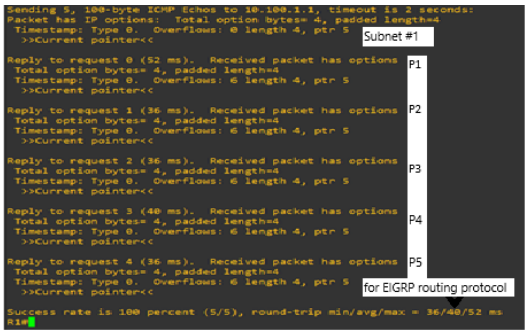


Figure. 4: The Ping utility executed on 10 routers utilizing the EIGRP protocol

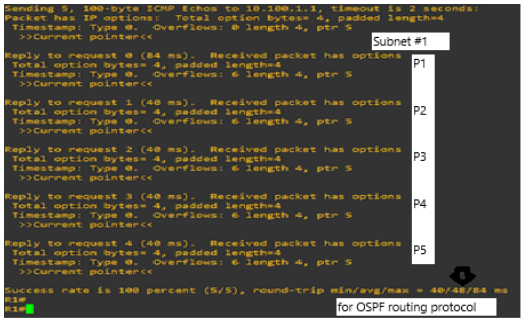


Figure 5: The Ping utility executed on 10 routers utilizing the OSPF protocol.

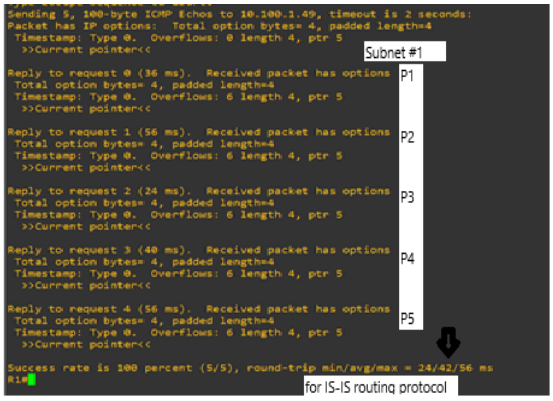


Figure 6: The Ping utility executed on 10 routers utilizing the IS-IS protocol.

The obtained results of the previous four Figures are measured using Ping utility using Internet Communication Message Protocol (ICMP) echoes. The round-trip is measured through packet traveling using different interior gateway routing protocol in each time using the same large-scale proposed network topology. As shown above, the results are different from one interior gateway routing protocol to other interior gateway routing protocol.

Figure 7 represents the delay variation, from the values in Table 3, so can see that for the 1st case EIGRP has the least delay, whereas the 2nd and the 3rd case OSPF has the least delay. For the 4th, 5th and 6th case EIGRP has the least delay. For the 7th and 8th case IS-IS has the least delay. For the 9th ,13th and 14th OSPF has the least value. So EIGRP has the least delay among the OSPF and IS-IS.

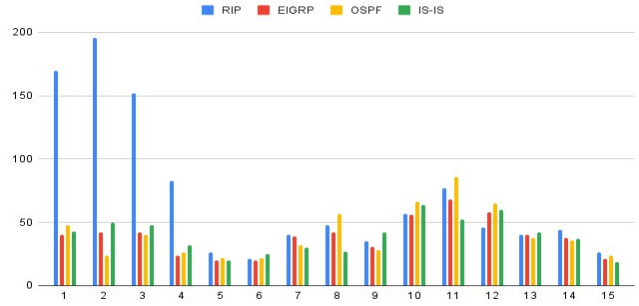


Figure 7: End-to-End Delay variation graph for RIP EIGRP, OSPF, and IS-IS.

The Table 4 below displays the average values of the end-to-end delay for RIP, EIGRP, OSPF, and IS-IS which has been computed by taking the average of the Table 3 averages for each protocol separately.

Table 4: Comparison of Average End to End Delay for RIP, EIGRP, OSPF and IS-IS.

S.No.	TYPE OF PROTOCOL	AVERAGE END TO END DELAY TIME (ms)
1	RIP	70.73
2	EIGRP	38.73
3	OSPF	40.93
4	IS-IS	39.33

The Figure 8 below represents the average end-to-end delay from the values in Table 4, and by analyzing Figure. 7, can see that EIGRP has the least average value of delay and does not differ much from IS-IS and OSPF. Figure 7 shows that OSPF has the worst average value of delay.

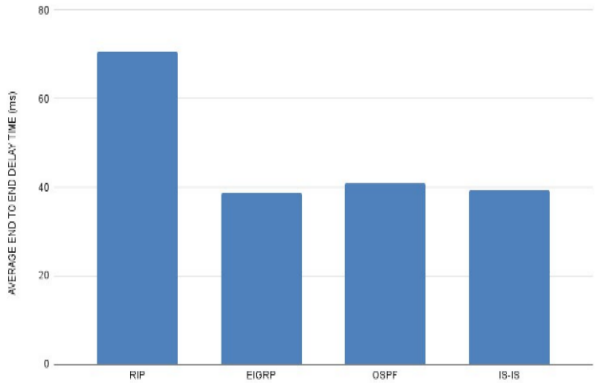


Figure 8: Delay Variation for RIP, EIGRP, OSPF and IS-IS.

7.2. *In depth analysis for Jitter Testing Results*

A) Proposed Method

In this section, to obtain the most accurate jitter result, a delay of 75 packets (fifteen (15) subnets multiplied by five (5) received messages using Ping utility using ICMP echoes) has been calculated for each protocol, and 60 jitter results were obtained for all interior gateway routing protocol By calculating the differences between the delays for the five packets, respectively, as shown in Tables 5, 6, 7, and 8 below.

An example of this is that the result of the first jitter is the difference between the delay for the first and second packets, and the result of the second jitter is the difference between the delay for the second and third packets, and so on. Let R represents the jitter result, i represents the number of received packet sequence within one subnet, and P represents packet delay as shown in (4).

$$R_i = |P_{i+1} - P_i| \tag{4}$$

The Table 9 below displays the average values of the in-depth analysis of jitter for RIP, EIGRP, OSPF, and IS-IS, which has been calculated by taking the average of 60 (the volume of the data packets is equal to fifteen (15) subnets multiply by four (4) different interior gateway routing protocols) values of jitter from the Tables 5, 6, 7, and 8 above for each protocol separately. Let R represent the Jitter result for each two send packets obtained in (4), i represent the iteration of measurement jitters in each subnet, and K represents the number of jitter results equal to 4 (i-th of 4) for each interior gateway routing protocol, and AVG is the average value of jitter measure calculated in each subnet iteration (i-th of 15).

$$AVG_{ith} = \frac{\sum_{i=1}^{K} R_i}{\mathrm{K}} \tag{5}$$

Table 5: In depth analysis JITTER result for RIP.

IP	End-to-End Delay					Jitter (Traditional Method)	Jitter (Ignored remained messages by traditional method of measuring Jitter calculation)			This Work Proposed Method
	Packet P1	Packet P2	Packet P3	Packet P4	Packet P5	Result R1=|P2-P1|	Result R2=|P3-P2|	Result R3=|P4-P3|	Result R4=|P5-P4|	AVG= (R1+R2+R3+R4)/4
10.100.1.1/30	176	172	168	172	164	4	4	4	8	5
10.100.1.5/30	184	308	144	168	176	124	164	24	8	80
10.100.1.9/30	139	136	144	172	172	3	8	28	0	9.75
10.100.1.13/30	72	88	84	84	88	16	4	0	4	6
10.100.1.17/30	40	16	28	12	36	24	12	16	24	19
10.100.1.21/30	40	24	20	16	8	16	4	4	8	8
10.100.1.25/30	44	40	36	48	36	4	4	12	12	8
10.100.1.29/30	40	80	24	72	28	40	56	48	44	47
10.100.1.33/30	36	44	36	36	24	8	8	0	12	7
10.100.1.37/30	40	80	56	52	60	40	24	4	8	19
10.100.1.41/30	72	96	56	76	88	24	40	20	12	24
10.100.1.45/30	76	52	56	16	32	24	4	40	16	21
10.100.1.49/30	40	36	44	40	40	4	8	4	0	4
10.100.1.53/30	48	76	20	36	40	28	56	16	4	26
10.100.1.57/30	40	32	20	20	20	8	12	0	0	5
Grand Average for overall large-scale enterprise network topology						24.46666667	All are used in new proposed method in this work.			19.25

Table 6: In depth analysis JITTER result for EIGRP

IP	End-to-End delay					Jitter (Traditional Method)	Jitter (Ignored remained messages by traditional method of measuring Jitter calculation)			This Work Proposed Method
	Packet P1	Packet P2	Packet P3	Packet P4	Packet P5	Result R1=|P2-P1|	Result R2=|P3-P2|	Result R3=|P4-P3|	Result R4=|P5-P4|	AVG= (R1+R2+R3+R4)/4
10.100.1.1/30	52	36	36	40	36	16	0	4	4	6
10.100.1.5/30	52	36	44	40	40	16	8	4	0	7
10.100.1.9/30	56	32	44	40	40	24	12	4	0	10
10.100.1.13/30	44	20	16	20	20	24	4	4	0	8
10.100.1.17/30	36	16	12	20	20	20	4	8	0	8
10.100.1.21/30	36	16	20	16	16	20	4	4	0	7
10.100.1.25/30	36	68	16	40	36	32	52	24	4	28
10.100.1.29/30	32	52	36	60	32	20	16	24	28	22
10.100.1.33/30	44	24	32	28	28	20	8	4	0	8
10.100.1.37/30	64	28	36	76	76	36	8	40	0	21
10.100.1.41/30	68	88	52	68	64	20	36	16	4	19
10.100.1.45/30	68	60	64	52	48	8	4	12	4	7
10.100.1.49/30	32	52	44	24	52	20	8	20	28	19
10.100.1.53/30	32	76	20	24	40	44	56	4	16	30
10.100.1.57/30	36	16	20	20	16	20	4	0	4	7
Grand Average for overall large-scale enterprise network topology						22.66666667	All are used in new proposed method in this work.			13.8

Table 7: In depth analysis JITTER result for OSPF

IP	End-to-End delay					Jitter (Traditional Method)	Jitter (Ignored remained messages by traditional method of measuring Jitter calculation)			This Work Proposed Method
	Packet P1	Packet P2	Packet P3	Packet P4	Packet P5	Result R1=|P2-P1|	Result R2=|P3-P2|	Result R3=|P4-P3|	Result R4=|P5-P4|	AVG= (R1+R2+R3+R4)/4
10.100.1.1/30	84	40	40	40	40	44	0	0	0	11
10.100.1.5/30	44	20	20	16	20	24	0	4	4	8
10.100.1.9/30	68	16	52	40	28	52	36	12	12	28
10.100.1.13/30	36	28	24	20	24	8	4	4	4	5
10.100.1.17/30	40	20	20	20	12	20	0	0	8	7
10.100.1.21/30	36	24	20	24	8	12	4	4	16	9
10.100.1.25/30	32	52	24	32	24	20	28	8	8	16
10.100.1.29/30	64	56	52	56	60	8	4	4	4	5
10.100.1.33/30	44	20	40	20	16	24	20	20	4	17
10.100.1.37/30	80	56	84	48	64	24	28	36	16	26
10.100.1.41/30	84	112	72	76	88	28	40	4	12	21
10.100.1.45/30	80	60	72	52	64	20	12	20	12	16
10.100.1.49/30	40	40	40	36	36	0	0	4	0	1
10.100.1.53/30	48	40	24	36	36	8	16	12	0	9
10.100.1.57/30	40	24	24	24	8	16	0	0	16	8
Grand Average for overall large-scale enterprise network topology						20.53333333	All are used in new proposed method in this work.			12.46666667

Table 8: In depth analysis JITTER result for IS-IS

IP	End-to-End delay					Jitter (Traditional Method)	Jitter (Ignored remained messages by traditional method of measuring Jitter calculation)			This Work Proposed Method
	Packet P1	Packet P2	Packet P3	Packet P4	Packet P5	Result R1=|P2-P1|	Result R2=|P3-P2|	Result R3=|P4-P3|	Result R4=|P5-P4|	AVG= (R1+R2+R3+R4)/4
10.100.1.1/30	76	40	32	40	28	36	8	8	12	16
10.100.1.5/30	92	36	40	40	44	56	4	0	4	16
10.100.1.9/30	88	28	40	44	40	60	12	4	4	20
10.100.1.13/30	40	68	16	16	20	28	52	0	4	21
10.100.1.17/30	36	32	24	8	4	4	8	16	4	8
10.100.1.21/30	40	20	20	24	24	20	0	4	0	6
10.100.1.25/30	12	56	44	16	24	44	12	28	8	23
10.100.1.29/30	24	12	44	36	20	12	32	8	16	17
10.100.1.33/30	68	24	40	44	36	44	16	4	8	18
10.100.1.37/30	80	76	68	8	92	4	8	60	84	39
10.100.1.41/30	32	60	40	76	52	28	20	36	24	27
10.100.1.45/30	72	72	56	52	48	0	16	4	4	6
10.100.1.49/30	36	56	24	40	56	20	32	16	16	21
10.100.1.53/30	56	16	80	20	16	40	64	60	4	42
10.100.1.57/30	1	40	16	8	32	39	24	8	24	23.75
Grand Average for overall large-scale enterprise network topology						29	All are used in new proposed method in this work.			20.25

In order to get the grand average (GAVG) of each jitter measured interior routing protocol in the proposed network topology that has fifteen (15) subnets the calculation in (6) has been used, using the obtained measured of average iteration values in (5) for each subnet (S of total of fifteen (15) in this experiment).

$$\text{GAVG} = \frac{\sum_{ith=1}^{S} AVG_{ith}}{S} \quad (6)$$

The Table 9 shows the calculated grand average of overall measured value for jitter in the proposed large-scale enterprise network topology, refer to Figure 2 above. These measured values are obtained using the proposed method that is presented in this work.

Table 9: Comparison of Average jitter for RIP, OSPF, EIGRP and IS-IS using proposed method

S.No.	TYPE OF PROTOCOL	JITTER GRAND AVERAGE (ms)
1	RIP	19.25
2	EIGRP	13.8
3	OSPF	12.46666
4	IS-IS	20.25

The Figure 9 below represents the grand average jitter from the values in Table IX, OSPF's routing protocol has the lowest average value of jitter compared to RIP protocols, EIGRP protocol, and IS-IS protocol. Especially IS-IS, the IS-IS protocol has the highest average value of jitter.

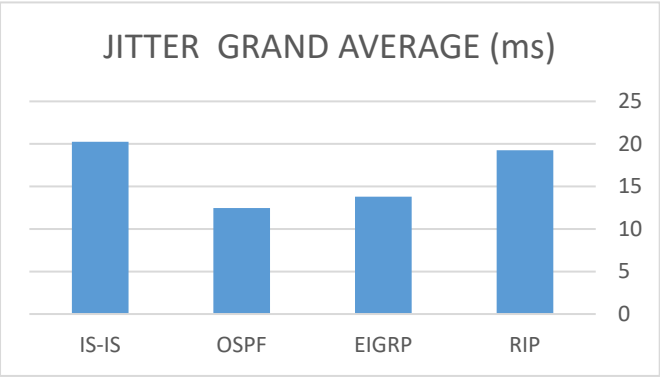


Figure. 9: Proposed Method for In Depth Jitter Variation Graph for RIP, EIGRP, OSPF and ISIS

B) Traditional Method to measure Jitter

This section also presents the classic technique of calculating Jitter, which is a well-known method for calculating the variance of only the first and second packets of replay to the requests messages using ICMP Echoes: [9], [10], [11] and [12].

Table 10 shows the calculated grand average of overall measured values for jitter in the proposed large-scale enterprise network topology, refer to Figure 2. These measured values are obtained using the traditional method of jitter metric, refer to Tables 5, 6, 7, and 8 above for each protocol separately.

Table 10: Comparison of Average jitter for RIP, OSPF, EIGRP and IS-IS using traditional method

S.No.	TYPE OF PROTOCOL	JITTER GRAND AVERAGE (ms)
1	RIP	24.46666667
2	EIGRP	22.66666667
3	OSPF	20.53333333
4	IS-IS	29

The Figure 10 below represents the grand average jitter from the values in Table 10, OSPF's routing protocol has the lowest average value of jitter compared to RIP protocols, EIGRP protocol, and IS-IS protocol. Especially IS-IS, the IS-IS protocol has the highest average value of jitter.

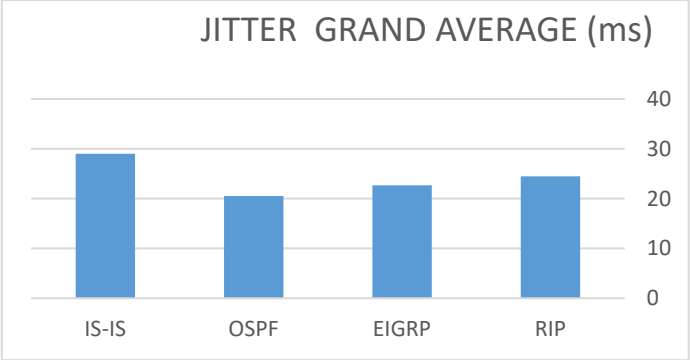


Figure 10: Traditional Method for Jitter Variation Graph for RIP, EIGRP, OSPF and ISIS

In order to get more clear of the results of both traditional mothed and proposed method, the comparing are shown in Figure 11 and in Table 11. The variance between both method is clear and indicate the difference that should be taken in account when measuring the Jitter metric especially in the large-scale enterprise network topology.

Table 11: Comparison of grand Average jitter for RIP, OSPF, EIGRP and IS-IS using traditional and proposed methods.

S. No.	Type of Protocol	Traditional Method (Jitter)	Proposed Method (Jitter)	The differences
1	RIP	24.46666667	19.25	5.216667
2	EIGRP	22.66666667	13.8	8.866667
3	OSPF	20.53333333	12.46666	8.066673
4	IS-IS	29	20.25	8.75

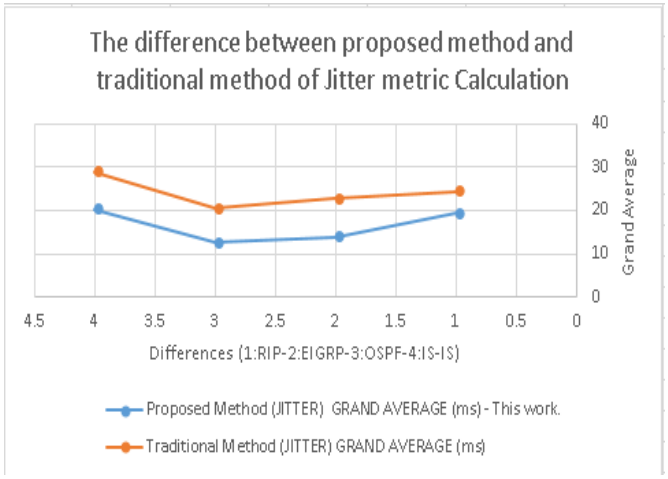


Figure 11: The veriation of Jitter metric using different measuring methods for In Depth Jitter Variation for RIP, EIGRP, OSPF and ISIS.

The Figure 11 above shows the differences that should be considered to pick the accuracy method to measure the Jitter in small, mid-sized, or large-scale network topology. The new method uses in-depth analysis technique using statistical method to improve the result of the measure using all reply messages to calculate the grand average when compare it to the traditional measuring method. The pseudo code for Jitter calculation using the proposed method in this work is illustrated program algorithm 1 below:

Algorithm 1: JITTER_PROPOSED_MEASURE.

Result: Grand_Average_Jitter_Measurement
Begin
REM %Initialization%;
REM SEND 100 Byte request ICMP Packet using PING Utility;
REM SEND 100 Byte request ICMP Packet using PING Utility.
REM % RECEIVE five (5) Messages by reply to requests%
REM %Let Packet represents as to receive messages time in milliseconds. %
REM % suppose the Received_Message_Time_ms array contains the time for the 5 received messages in milliseconds using ICMP Echo. %
double Traveled_Packet[1..5];
int Result [1..4] = 0;
double Received_Message_Time_ms[1..5] ← get time in milliseconds;
double SUM = 0.0;
double AVERAGE [1..15] = 0.0;
double AVERAGE_SUM = 0.0;
int K= 4,
int Number_of_Request_Messages = 5;
int Number_of_Subnets = 15;
For (**int** i:=1 to Number_of_Request_Messages)
    Traveled_Packet[i] =
    Received_Message_Time_ms[i];
Next i
REM %Calculate the variance between received messages for each subnet in the large-scale enterprise network topology. %
For (**int** S=1 to Number_of_Subnets)
For (**int** i=1 to Number_of_Request_Messages – 1)
    Result[i] = abs(Packet[i+1]-Packet[i]);
    SUM = SUM+Result[i];
Next i
  AVERAG[S] = SUM/K;
   AVERAGE_SUM = (AVERAGE_SUM +
   AVERAG[S]);
Next S
GRAND_AVERAG_JITTER_MEASURE = (AVERAGE_SUM/ Number_of_Subnets);
Print GRAND_AVERAG_JITTER_MEASURE;
END

7.3. *Convergence Time Testing Results*

When a route becomes unavailable, the router requires a certain amount of time to notify other routers of the failure and determine the best alternative path [13], [14], [15] and [16].

This process is commonly referred to as the convergence of time. In our experiment, packets of 100-datagram size and a 2-second timeout were transmitted from the source to the destination with a repeat count of 500. While sending these packets, we intentionally disrupted one of the routes and observed the subsequent loss of packets until the new path information was fully propagated to all routers. To determine the convergence time, we calculated the number of lost packets and multiplied it by the 2-second timeout. This process was repeated five times, and the average was calculated. Refer to Figures 12, 13, and 14 below.

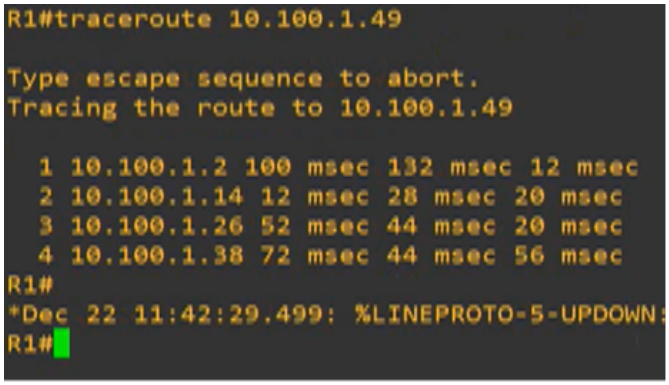


Figure 12: Traceroute for 10.100.1.49 Before Failure

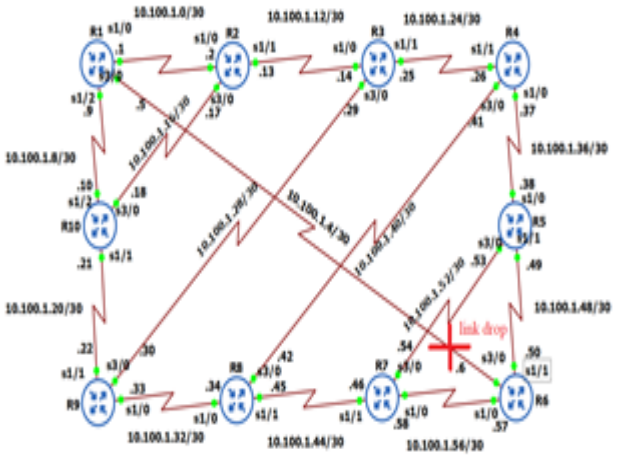


Figure 13: Link Failure for 10.100.1.49.

```
R1#traceroute 10.100.1.49

Type escape sequence to abort.
Tracing the route to 10.100.1.49

  1 10.100.1.6 92 msec 88 msec 88 msec
  2 10.100.1.49 88 msec 164 msec 104 msec
R1#
```

Figure. 14: Traceroute for 10.100.1.49 After Failure.

Table 12: Convergence Time of RIP.

S.No.	Packets received from 500 packets	Packet lost	Convergence time (s)	Average convergence time (s)
1	468	32	64	63.6
2	464	36	72	
3	471	29	58	
4	474	26	52	
5	464	36	72	

Table 13: Convergence Time of EIGRP

S.No.	Packets received from 500 packets	Packet lost	Convergence time (s)	Average convergence time (s)
1	489	11	22	20.8
2	488	12	24	
3	490	10	20	
4	489	11	22	
5	492	8	16	

Table 14: Convergence time of ospf

S.No.	Packets received from 500 packets	Packet lost	Convergence time (s)	Average convergence time (s)
1	481	19	38	40.8
2	480	20	40	
3	479	21	42	
4	479	21	42	
5	479	21	42	

Table 15: Convergence Time of IS-IS.

S.No.	Packets received from 500 packets	Packet lost	Convergence time (s)	Average convergence time (s)
1	486	14	28	30
2	484	16	32	
3	485	15	30	
4	485	15	30	
5	485	15	30	

Tables 12, 13, 14, and 15 above represent the convergence times and the average convergence times for each protocol. Figures 15, 16, 17, and 18 below represent samples of results.

```
R1#ping 10.100.1.49 repeat 500

Type escape sequence to abort.
Sending 500, 100-byte ICMP Echos to 10.100.1.49, timeout is 2 seconds:
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

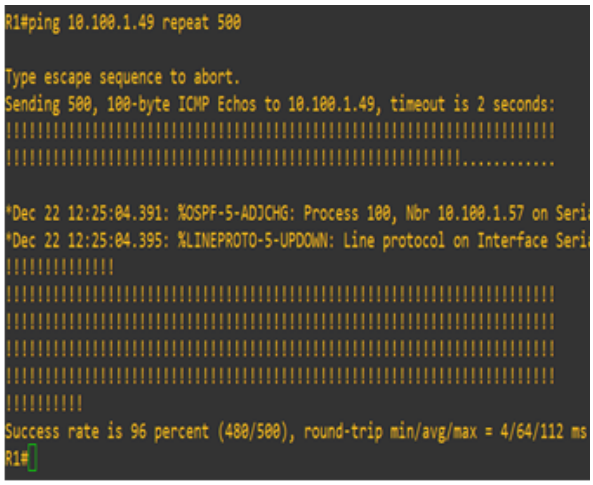


Figure 17: OSPF Configuration When R6's Port s3/0 is Blocked.

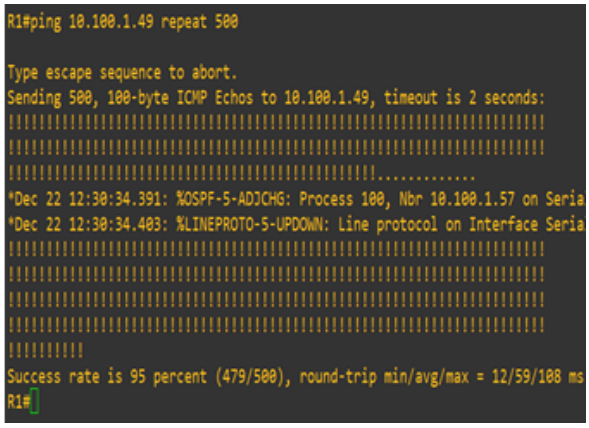


Figure 18 : IS-IS Configuration When R6's Port s3/0 is Blocked.

The Figure 19 below represents the average of convergence time from the values in Tables 12, 13, 14, and 15 above, and by analyzing Figure 19, so can see that EIGRP has the least average value of convergence time and the OSPF has the worst average value of convergence time.

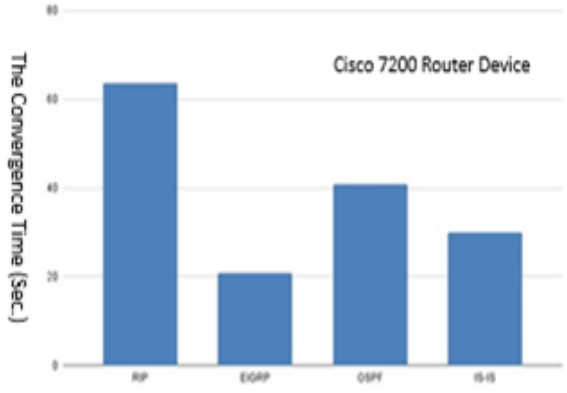


Figure 19: Convergence Time Variation Graph for RIP, EIGRP, OSPF and IS-IS.

As explained in the beginning of the sub-section "7.3", conversion time is calculated after the failure of one or more network router device in the network topology and the least measured value is the best recoverable protocol. The obtained result from the experiment indicate that the EIGP is more suitable for large-scale enterprise network topology.

8. Discussion

Many research papers have been presented the evaluation of the interior gateway routing protocols in different network topologies, each topology has different characteristics like the number of routing paths that used in their experiments, number of network devices, topology type as small; mid-sized; or large-scale network, number of subnets used in their experiments, internet protocol versions that used, default cost variation, route table size, dropped traffic and so on:[16][17][18] and [19].

Also there are different techniques have been used to measure the metric of these interior gateway routing protocols regardless of the network topology especially when measure the jitter to evaluate the performance of the network: [12] [19] and [20]. Here, in this work, the evaluation of these interior gateway protocols has been conducted in a large-scale network topology consist of 30-routers with tri-connected interfaces. A new method to calculate the Jitter has been presented in detail with its pseudo code and then compared with the simple technique method of measuring the Jitter in the computer networks. The result has shown there is a big different in term of the accuracy of both calculation methods, but both the jitter measurement methods have the same curve pattern.

9. Conclusion

This study compared the performance of RIP, EIGRP, OSPF, and IS-IS inner gateway protocols in terms of convergence time, end-to-end delay, and jitter for fifteen subnets in proposed network topologies. The results show that EIGRP outperformed other routing protocols in terms of latency and convergence time, although OSPF outperformed them in terms of jitter. In addition, IS-IS outperformed OSPF in terms of convergence time. In contrast, the results showed that RIP protocols had the poorest delay and convergence time performance. Finally in this work, it has introduced a new method for measuring jitter. The formula was developed to calculate network jitter with greater accuracy than the traditional, convention, method using the grand average calculations. Additionally, the suggested method may be replicated to model various network topology sizes, and the new approach can be contrasted with other methods to improve the speed and precision of jitter measurement in the field of routing networks. This would develop and improve new monitoring tools to measure jitter networks more accurately and quickly.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Usman, E.E. Oberafo, M.A. Abubakar, T. Aminu, A. Modibbo, S. Thomas, "Review of interior gateway routing protocols," in *2019*

15th International Conference on Electronics, Computer and Computation, ICECCO 2019, 2019, doi:10.1109/ICECCO48375.2019.9043204.

[2] (PDF) Performance Comparison of Routing Protocols | yousef aburawi - Academia.edu, Oct. 2024.

[3] A.G. Biradar, "A Comparative Study on Routing Protocols: RIP, OSPF and EIGRP and Their Analysis Using GNS-3," *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering, ICRAIE 2020 - Proceeding*, 2020, doi:10.1109/ICRAIE51050.2020.9358327.

[4] C. Filsfils, A. Bashandy, H. Gredler, B. Decraene, "IS-IS Extensions for Segment Routing," 2019, doi:10.17487/RFC8667.

[5] I. Joseph Okonkwo, I. Douglas Emmanuel, "Comparative Study of EIGRP and OSPF Protocols based on Network Convergence," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020.

[6] D.R. Prehanto, A.D. Indriyanti, G.S. Permadi, "Performance analysis routing protocol between RIPv2 and EIGRP with termination test on full mesh topology," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, 354–361, 2021, doi:10.11591/IJEECS.V23.I1.PP354-361.

[7] M. Athira, L. Abrahami, R.G. Sangeetha, "Study on network performance of interior gateway protocols — RIP, EIGRP and OSPF," *2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, 344–348, 2017, doi:10.1109/ICNETS2.2017.8067958.

[8] A. Manzoor, M. Hussain, S. Mehrban, "Performance Analysis and Route Optimization: Redistribution between EIGRP, OSPF & BGP Routing Protocols," *Computer Standards & Interfaces*, vol. 68, 103391, 2020, doi:10.1016/J.CSI.2019.103391.

[9] (PDF) Performance Analysis of Interior Gateway Protocols, https://www.researchgate.net/publication/303812574_Performance_Analysis_of_Interior_Gateway_Protocols, Oct. 2024.

[10] H. Karna, V. Baggan, A.K. Sahoo, P.K. Sarangi, "Performance Analysis of Interior Gateway Protocols (IGPs) using GNS-3," *Proceedings of the 2019 8th International Conference on System Modeling and Advancement in Research Trends, SMART 2019*, 204–209, 2020, doi:10.1109/SMART46866.2019.9117308.

[11] What Is Jitter, and How Does It Affect Your Internet Connection?, https://www.howtogeek.com/824032/what-is-jitter/, Oct. 2024.

[12] How to Measure Jitter - Obkio, https://obkio.com/blog/how-to-measure-jitter/, Oct. 2024.

[13] IMPROVING THE PERFORMANCE OF AD HOC NETWORK USING ENHANCED INTERIOR GATEWAY ROUTING TECHNIQUE - IJSER Journal Publication, https://www.ijser.org/onlineResearchPaperViewer.aspx?IMPROVING_THE_PERFORMANCE_OF_AD_HOC_NETWORK_USING_ENHANCED_INTERIOR_GATEWAY_ROUTING_TECHNIQUE.pdf, Oct. 2024.

[14] N. Kumari, D.T. Raj, "OSPF Metric Convergence and Variation Analysis During Redistribution with Routing Information Protocol," *International Research Journal on Advanced Engineering and Management (IRJAEM)*, vol. 2, no. 06, 1985–1991, 2024, doi:10.47392/IRJAEM.2024.0293.

[15] E.C. Obioma, E.C. Amadi, "Analysis of Optimal Routing Protocol for an Inter-Campus Private Cloud Network System," *International Journal of Modern Developments in Engineering and Science*, vol. 3, no. 2, 8–12, 2024, doi:10.5281/ZENODO.12643596.

[16] K. Shahid, S.N. Ahmad, S.T.H. Rizvi, "Optimizing Network Performance: A Comparative Analysis of EIGRP, OSPF, and BGP in IPv6-Based Load-Sharing and Link-Failover Systems," *Future Internet*, vol. 16, no. 9, 2024, doi:10.3390/FI16090339.

[17] N. Kumari, A. Sharma, "OSPF Default Cost Variation for Type-1 and Type-2 External Links during Redistribution," *International Journal of Computer Applications*, vol. 184, no. 28, 8–14, 2022, doi:10.5120/IJCA2022922345.

[18] M.E. Mustafa, ; Ahmed, A. Abaker, "International Journal of Computer Science and Mobile Computing Evaluate OSPF and EIGRP Based on Route Table Size and Dropped Traffic using Different Topologies (Star, Loop)," *International Journal of Computer Science and Mobile Computing*, vol. 13, 6–17, 2024, doi:10.47760/ijcsmc.2024.v13i04.002.

[19] D. Turlykozhayeva, S. Temesheva, N. Ussipov, A. Bolysbay, A. Akhmetali, S. Akhtanov, X. Tang, "Experimental Performance Comparison of Proactive Routing Protocols in Wireless Mesh Network Using Raspberry Pi 4," *Telecom 2024, Vol. 5, Pages 1008-1020*, vol. 5, no. 4, 1008–1020, 2024, doi:10.3390/TELECOM5040051.

[20] Performance Analysis of Routing Protocols RIP, EIGRP, OSPF and IGRP using Networks connector, 2020, doi:10.4108/eai.28-6-2020.2298167.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



Saleh Hussein Al-Awami has done his bachelor's degree from University of Benghazi formerly named Garyounis University in 1996. He has done his master's degree in Computer Sciences - Artificial Intelligent from The Libyan Academy for Postgraduate Studies in Benghazi-Libya in 2010. His researches interested in Artificial Intelligent, Internet of Things (IoT), Computer Networks, Cloud Computing, Software Defined Networks, Computer Network Infrastructures Security, Fintech infrastructure and application services and System Design and Methodologies. He has many scientific papers and journals publications related to his researches. Saleh has published scientific papers in many conferences as in United States (Las-Vegas), India (Mumbai), Saudi-Arabia(Riyadh), Jordan, and in Libya. He has hold some attendance professional certificates from several international training centers in United Kingdom (UK-London) and Egypt (Cairo) holding CompTIA N+, Cisco Certified Network Associate (CCNA, 2002), Microsoft SQL databases, Microsoft Exchange Server (E-mail Systems administrator since 2007 edition), and Microsoft Server operating systems since 2003 edition. Also, he has hold some professional certificates as Microsoft network administrator, Novel network services administrator, and recently Oracle Cloud Infrastructure 2024 Generative AI certificated Professional.



Emad Awadh Ben Srity completed his Bachelor's degree in Computer Networks and Data Communications at the University of Benghazi in 2023. He is passionate about computer network research, network design, and analysis. Emad has published a scientific paper presented at a Libyan conference and received training in computer systems at Arabian Gulf Oil Company and ISP Smart Link Company. He worked successfully at an ISP company for about a year and also got an opportunity with one of the companies working on the maintenance of the Faculty of Science at the University of Benghazi. He worked as a supervisor of the Physics and Geology departments in terms of electricity and networks. He remained a site manager at the Faculty of Law in terms of electricity and networks, and worked on designing network maps, including the distribution of fire alarm points, motion sensors, access points, data, and telephone lines.



Ali Tahir Abu Raas has done his bachelor's degree from Computer Networks and Data Communications - University of Benghazi in 2023. He has interested in computer networks researches, network design and analysis. He received training in computer systems at Arabian Gulf Oil Company and ISP Company. Ali published a scientific paper in 2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA).

Received: 26 August, 2024, Revised: 09 October, 2024, Accepted: 10 October, 2024, Online: 29 October, 2024

DOI: https://doi.org/10.55708/js0310006

Biclustering Results Visualization of Gene Expression Data: A Review

Haithem Aouabed* [1], **Mourad Elloumi** [2], **Fahad Algarni** [2]

[1]University of Sfax, Department of Computer Sciences, Faculty of Economic Sciences and Management, Sfax, 3018, Tunisia

[2]University of Bisha, Department of Computer Sciences and artificial intelligence, College of Computing and Information Technology, Bisha, 67714, Saudi Arabia

*Corresponding author: Haithem Aouabed, Mahdia road km 19 Sfax, +21626399303 &haithem.abdi@gmail.com

ABSTRACT: Biclustering is a non-supervised data mining method used to analyze gene expression data by identifying groups of genes that exhibit similar patterns across specific groups of conditions. Discovering these co-expressed genes (called biclusters) can aid in understanding gene interactions in various biological contexts. Biclustering is characterized by its bi-dimensional nature, grouping both genes and conditions in the same bicluster and its overlapping property, allowing genes to belong to multiple biclusters. Biclustering algorithms often produce a large number of overlapping biclusters. Visualizing these results is not a straightforward task due to the specific characteristics of biclusters. In fact, biclustering results visualization is a crucial process to infer patterns from the expression data. In this paper, we explore the various techniques for visualizing multiple biclusters simultaneously and we evaluate them in order to help biologists to better choose their appropriate visualization techniques.

KEYWORDS: Biclustering algorithms, Biclusters, Overlaps, Visualization, Visualization techniques

1. Introduction

Gene expression profiles, generated by advanced high-throughput technologies like microarrays, are depicted in a matrix format where rows correspond to genes, columns to experimental conditions and each matrix entry to the expression level of a gene under a specific condition. Clustering has been the primary technique for analyzing such voluminous genomic data, focusing on grouping genes (rows) that show similar expression across all conditions (columns), as noted by [1]. Traditional clustering methods including hierarchical clustering [2] and k-means clustering [3] have proven effective in gene expression analysis. However, to glean novel insights from biological data such as identifying genes associated with cancer progression, determining functions of unknown genes or developing new treatment approaches, it's essential to perform clustering across both dimensions: genes and conditions. Indeed, the field of biclustering or co-clustering has emerged as a valuable tool in genomic data analysis. This machine learning technique identifies groups of biological entities, such as genes, that display comparable behaviour under specific conditions. It is first used to analyze gene expression data in 2000 by [4]. Biclustering differs from traditional clustering in two key theoretical aspects: bi-dimensionality which involves grouping genes and conditions together and overlap which permits genes to be part of multiple biclusters at the same time. In [5], the author conducted a comprehensive review of various biclustering algorithms, categorizing them based on their search methodologies.

Visualizing biclustering output allows for the identification of co-regulated gene clusters and experimental conditions with similar gene expression profiles. In fact, examining biclustering results visually provides a deeper understanding of the underlying relationships and trends within the expression data [6]. Nevertheless, due to the unique attributes of biclustering which are bi-dimensionality and the potential for overlaps, the representation of gene expression data often results in numerous intersecting biclusters. These are challenging to display comprehensively in an informative way in a single visual representation. Indeed, encapsulating the results of biclustering into a single, coherent visual format is far from straightforward. Finding novel insights from vast, intricate multi-dimensional datasets necessitates an effective synergy of data processing algorithms and the power of interactive visualization tools [7-11]. Such a blend has been effectively applied to biological datasets, exemplified by the biclustering of gene expression data [12]. While heatmaps [1] and parallel coordinates [13]

remain the go-to methods for visualizing individual biclusters [14-16], the real challenge emerges when attempting to concurrently visualize multiple biclusters on a single screen for bioinformaticians and analysts [17], [18].

In this review, we conduct an examination of the visualization methods applied to biclustering outcomes derived from gene expression data [17]. We focus on biclustering results visualization techniques that can show more than one bicluster in the same screen. The structure of the review is outlined as follows: First, we provide an overview of the biclustering concept as it applies to gene expression data. Second, we make a survey on the current methods available for visualizing several biclusters concurrently. Then, we evaluate these methods according to a set of predefined criteria. Next, we demonstrate practical applications of these methods through various tools and discuss the datasets employed for their validation. Finally, we offer our conclusion.

2. Biclustering of gene expression data concept

We start by giving a definition of the biclustering concept.

2.1. Definition

A *bicluster* is a group of genes that exhibit consistent patterns of expression across a specific set of conditions. These genes have similar expression levels or follow identical trends within these conditions [19]. We note that biclusters can *overlap*, meaning that individual genes or conditions may be part of multiple biclusters at the same time.

Formally, a bicluster can be defined as follows: Let I={1, 2, . . . , n} be a set of indices of n genes, J={1, 2, . . . ,m} be a set of indices of m conditions and $M\ (I, J)$ be a data matrix associated with I and J. A bicluster associated with the data matrix $M\ (I, J)$ is a couple (I', J') such that $I' \subseteq I$ and $J' \subseteq J$.

The *biclustering problem* can be formulated as follows: Given a data matrix M, construct a group of biclusters B_{opt} associated with M such that:

$$f(B_{opt}) = \max_{B \in BC(M)} f(B), \quad (1)$$

where f is a function that evaluates the quality, or coherence, of a bicluster group and $BC(M)$ represents the set of all potential bicluster groups associated with M[20], [21]. Biclustering is an NP-hard problem [4], [22]. In fact, NP-hard problems (i.e., non-deterministic polynomial time problems) are a class of challenging problems that are often considered intractable, meaning that there is no known efficient algorithm for solving them a priory. The combinatorial nature of the search space and the multiple optimization criteria involved make biclustering of gene expression data an NP-hard problem.

2.2. Groups of biclusters

Bicluster groups can be classified as follows [22]. See Figure 1:

- Particular bicluster (a).
- Bicluster groups with unique rows and columns (b).
- Checkerboard bicluster groups without overlap (c).
- Bicluster groups with unique rows (d).
- Bicluster groups with unique columns (e).
- Tree-structured bicluster groups without overlap (f).
- Non-overlapping bicluster groups without exclusive membership (rows or columns) (g).
- Hierarchical bicluster groups with overlap (h).
- Randomly placed overlapping group of biclusters (i).

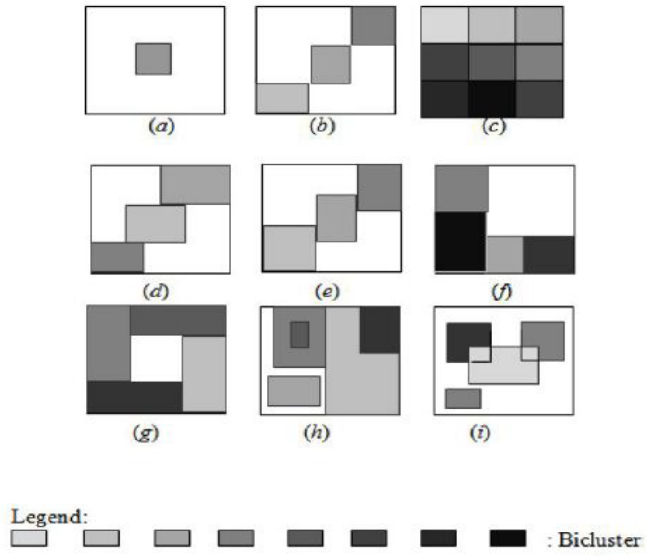


Figure 1: Types of bicluster groups [20]

We focus our review on how to visualize more than one bicluster with overlaps in the same screen (Figure 1(h) and Figure 1(i)).

2.3. Biclustering search methods

Biclustering is a computationally complex problem, often classified as NP-hard [4], [22]. As a result, heuristic approaches are typically employed to find approximate solutions. Given the variety of biclustering algorithms based on different search strategies, the following categorization can be identified [22], [23]:

- *Iterative row/column clustering*: This approach is a straightforward method that involves applying clustering algorithms to both the rows and columns of the expression matrix and then combining the results to identify biclusters. It is also known as two-way clustering [24], [25] or conjugated clustering [26]. This approach inherits the same benefits and drawbacks of clustering algorithms. Examples of algorithms that follow this approach include ITWC (Interrelated Two-Way Clustering) [25], CTWC (Coupled Two-Way Clustering) [24] and DCC (Double Conjugated Clustering) [26].

- *Divide and conquer*: This approach begins with a single bicluster encompassing the entire data matrix. It then recursively divides this matrix into two submatrices, creating two new biclusters. This process continues until a predetermined number of biclusters are generated that meet specific criteria. By breaking down the problem into smaller subproblems, this approach aims to accelerate the search for solutions. While this method is known for its speed, it can potentially ignore valuable biclusters if they are divided before being identified [21]. Examples of algorithms that follow this approach include the Hartigan biclustering algorithm [27] and Bimax[28].
- *Greedy iterative search*: This approach builds a solution incrementally using a specified quality measure. In the context of biclustering, at each step, submatrices of the data matrix are constructed by adding/removing rows or columns to maximize/minimize a particular function. This process continues until no further modifications can be made to any submatrix [20]. This approach shares the same strengths and weaknesses as the divide-and-conquer method. While it may make suboptimal choices and miss good biclusters, it can be very fast [21]. Examples of algorithms that follow this approach include CCA [4], OPSM [29], xMOTIFs[30], ISA [31], MSSRCC [32], QUBIC [33], COALESCE [34], CPB [35] and LAS [36].
- *Exhaustive bicluster enumeration*: This approach exhaustively explores all potential bicluster groups to identify the optimal solution that maximizes a specific evaluation function. Despite the capability of finding the best results, this approach is computationally expensive (i.e., time consuming). To alleviate this, biclustering algorithms often incorporate restrictions on the size or number of biclusters or employ pre- and post-filtering techniques [28]. Examples of algorithms that follow this approach include SAMBA [37], BiBit[38] and DeBi[39].
- *Distribution parameter identification*: This approach employs a statistical model to estimate distribution parameters and generate data by iteratively minimizing a specific criterion. Algorithms that follow this approach are capable of identifying the optimal biclusters, if they exist. However, due to their high computational complexity, they are often limited to analyzing biclusters with a specific size [21]. Examples of algorithms that follow this approach include Plaid model [40], Spectral biclustering[41], BBC [42] and FABIA [43].

Table 1 is a summary of biclustering search methods algorithms describing their characteristics.

Table 1: Evaluation of biclustering search methods

Biclutering search method	Algorithms	Advantages	Disadvantages
Iterative row/column clustering	ITWC [25] CTWC [24] DCC [26]	Find good results (i.e., clusters) Very fast	Sensitivity to noise datasets Scalability issues to large datasets
Divide and conquer	Hartigan algorithm [27] Bimax[28]	Very fast	Ignore good biclusters
Greedy iterative search	CCA [4] OPSM [29] xMOTIFs[30] ISA [31] MSSRCC [32] QUBIC [33] COALESCE [34] CPB [35] LAS [36]	Very fast	Ignore good biclusters
Exhaustive bicluster enumeration	SAMBA [37] BiBit[38] DeBi[39]	Find best solutions (i.e., biclusters)	Very Slow Time consuming
Distribution parameter identification	Plaid model [40] Spectral biclustering[41] BBC [42] FABIA [43]	Find best solutions (i.e., biclusters)	High complexity Time consuming

Most of these biclustering algorithms already mentioned in the previous subsection generated generally big size biclusters as a result. In the next section, we present the techniques used to visualize biclustering results of gene expression data [17].

3. Review of previous research

3.1. Heatmaps visualization

A heatmap serves as a bi-dimensional graphical representation that illustrates data values within a matrix structure. For gene expression data, the x-axis is allocated for conditions (columns) and the y-axis for genes (rows). Each matrix element a_{ij} denoting the expression level of the i^{th} gene under the j^{th} condition. It is depicted as a colored square (pixel) with the color intensity corresponding to a predefined scale. Typically, shades of green, red and black colors are chosen to align with the standard fluorescent dyes used in DNA microarrays: green signifies reduced expression or down-regulation, red indicates elevated expression or up-regulation and black represents a neutral expression level. To visualize a bicluster, its associated rows and columns are repositioned, generally to the top-left corner of the matrix [44]. Visualization techniques such as reordering or duplication are employed to display multiple biclusters within a single view. An example of heatmaps displaying biclusters is depicted in Figure 2.

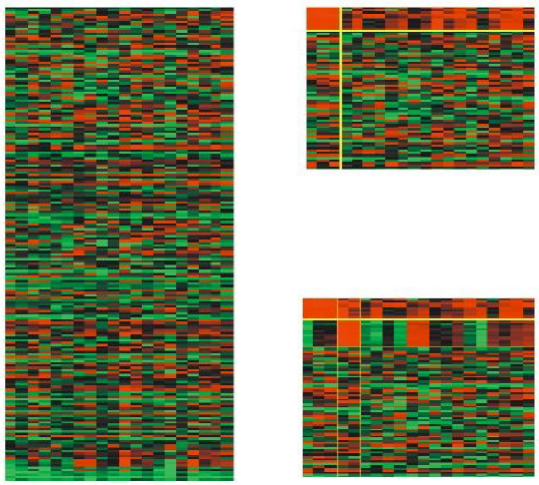
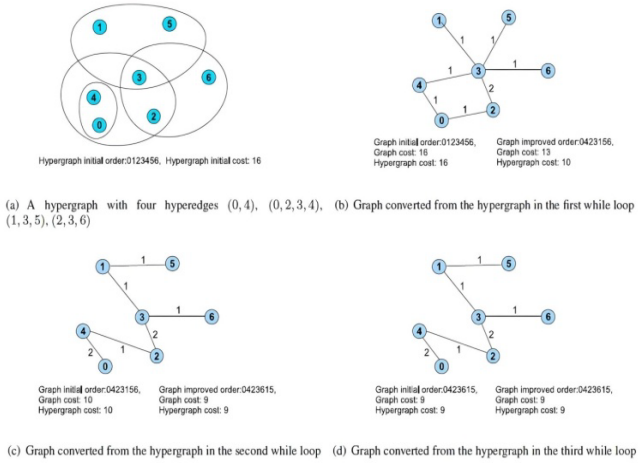


Figure 2: Heatmap representation of gene expression data (on the left). A bicluster at the upper left corner (top right). Two biclusters on the diagonal of the matrix (bottom right) [45]

3.1.1. Reordering techniques

To simultaneously visualize multiple biclusters, reordering is a viable strategy for heatmap representations. The literature presents various algorithms for this purpose.

In [46], the author proposed a heuristic iterative method that approaches the visualization of overlapping biclusters as an optimization challenge. This method introduces a reordering technique that draws parallels with the *hypergraph vertex ordering dilemma*, an extension of the classic *minimal linear arrangement* or *graph ordering* problem. Initially, the heatmap matrix is transformed into a hypergraph which is then converted into a weighted undirected graph following a starting order that aligns with one of three predetermined configurations: a linear path, a singular loop or multiple loops. Then, the minimum linear arrangement problem (i.e., *MinLA*) is employed on the newly formed graph to discover an improved order. Next, the hypergraph is transformed into a different graph using the newly determined order. This iterative process continues until a satisfactory order is achieved or no further enhancements are observed. An illustrative example of this algorithm in action is presented in Figure 3.



(a) A hypergraph with four hyperedges $(0,4)$, $(0,2,3,4)$, $(1,3,5)$, $(2,3,6)$ (b) Graph converted from the hypergraph in the first while loop (c) Graph converted from the hypergraph in the second while loop (d) Graph converted from the hypergraph in the third while loop

Figure 3: A practical demonstration of the reordering algorithm [46]

In [47], the algorithm presented aims to optimize the layout to enhance the visualization of the largest contiguous sections of biclusters from a gene expression matrix. Initially, the data is depicted as a *binary matrix* with rows representing genes or conditions and columns representing biclusters. The reordering approach is independently applied to both rows and columns to improve the visual quality of the biclusters. This optimization process counts four stages: The first stage named *simplify* eliminates redundant rows to reduce the problem's complexity. The second stage named *prearrange* seeks an optimal starting point for optimization by sequentially adding rows to a new order, ensuring each is positioned ideally. The third stage named *arrange* is the core of the algorithm where it aims to maximize an alignment score using a greedy strategy that repositions parts of a bicluster and permutes its constituent elements (genes or conditions) for better alignment. The final stage named *complexity* reintroduces the previously excluded rows into their new positions, thereby restoring the original problem's scale. An illustration of this technique's workflow is provided in Figure 4.

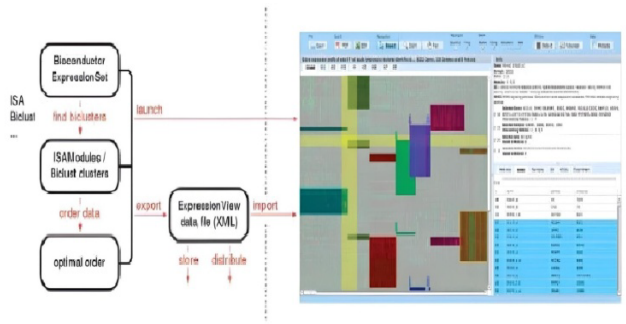


Figure 4: Different tasks of the proposed technique. Analysis part (on the left). Heatmap visualization (on the right) [47].

3.1.2. *Duplication techniques*

In certain instances within heatmap visualizations, reordering techniques can't display all biclusters adequately. It becomes necessary to replicate rows and columns to present the biclusters as continuous segments within a single heatmap. This duplication strategy has been proposed in various studies to enhance the clarity and continuity of bicluster representation.

The algorithm introduced by [48] aims to visualize biclusters and their overlaps as continuous regions within a single heatmap. Its core concept involves duplicating rows and columns to accurately depict overlapping biclusters. This approach is influenced by the *hypergraph superstring challenge* which refers to the physical mapping of genomes, as investigated by [49]. The algorithm presents a technique to limit the duplication of rows and columns as much as possible. This technique is executed separately on both rows and columns. It employs a data structure known as a *PQ tree*[50] which is helpful to arrange all potential columns to be adjacent, duplicating them if necessary, to form contiguous biclusters. Additionally, it utilizes a sequence of *REDUCE* operations that hierarchically organize the rows, thereby enhancing the overall quality of the visualization. An illustrative example employing two distinct expression matrices is demonstrated in Figure 5.

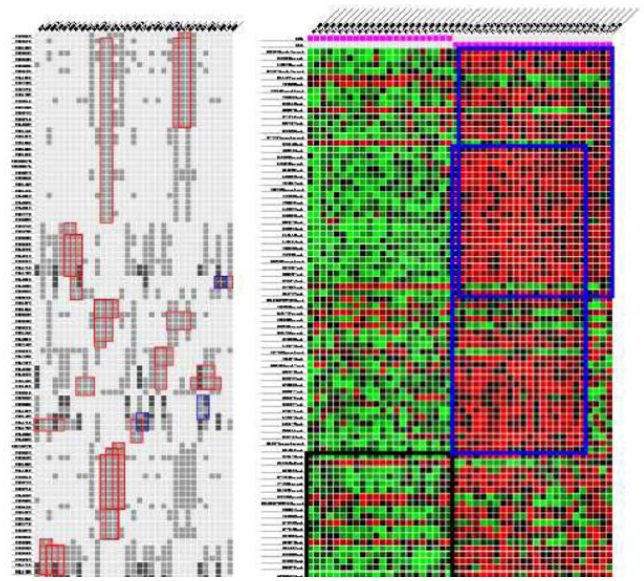


Figure 5: A number of biclusters were depicted [48]. Red rectangles (on the left) and blue rectangles (on the right)

In [51], the author developed a biclustering layout algorithm alongside an interactive visualization interface for illustrating multiple biclusters. The initial phase of their algorithm involves translating the heatmap into grayscale values through linear interpolation, spanning from the minimum to the maximum values within the data matrix. Next, each bicluster is designated with a unique color. When selected, biclusters are highlighted in a *semi-transparent yellow hue* which merges additively in areas of overlap, although users retain the option to customize their colors. To enable analysts to selectively visualize biclusters in an adjacent manner, the algorithm incorporates both reordering and duplication strategies for rows and/or columns. This interactive feature significantly reduces the occurrence of duplicates and marginally enhances the method's scalability. Figure 6 illustrates varied heatmap visualizations drawn from three distinct datasets containing multiple biclusters.

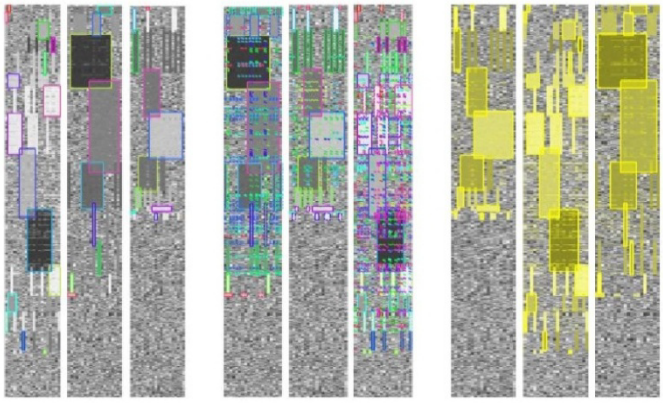


Figure 6: Bicluster results visualization. Each bicluster is depicted by its main rectangle (on the left). All biclusters are shown (in the middle). Representation with emphasized biclusters (on the right) [51]

Despite being the most common technique for visualizing single biclusters, heatmaps have limitations in terms of geometry especially when displaying biclusters with high levels of overlap.

3.2. *Parallel coordinates visualization*

Parallel coordinates are employed as a visualization method for representing complex, high-dimensional data sets. In this technique, each dimension is associated with a vertical line and individual data points are connected across these lines to form a *polyline* that reflects their multi-dimensional values. This approach has been adapted for the visualization of gene expression data as well. To depict gene profiles within an m-dimensional framework, *m* parallel and equidistant vertical lines are drawn, each symbolizing a different experimental condition. The gene profiles are then plotted as polylines across these lines, with the position of each point on a line corresponding to the gene's expression level under that particular condition. An example of this visualization technique, using parallel coordinates, is provided in Figure 7.

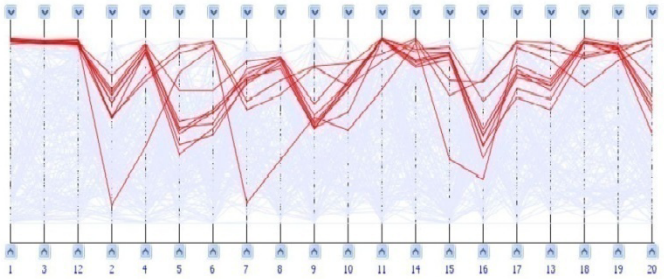


Figure 7: Parallel coordinates visualization. Polylines of significant genes highlighted in red [45]

In [51], the author implemented a series of transformations to their heatmap data representation in order to facilitate the simultaneous visualization of multiple biclusters using parallel coordinates. In this adaptation, the matrix of rows is represented as lines within the parallel coordinates framework. The vertical axes are positioned to correspond with the columns from the heatmap. To depict the conditions associated with a bicluster, the method computes the *mean vertical location* of all lines within a bicluster, establishing reference points known as *centroids*. These lines are then adjusted to intersect at the centroids. Next, the biclusters are rendered in a semi-transparent black hue. The color scheme utilized for the heatmap biclusters is replicated in the parallel coordinates display. To reduce visual clutter, lines not part of the highlighted biclusters can be dimmed by the user. An illustration of this parallel coordinates visualization for two biclusters is presented in Figure 8.

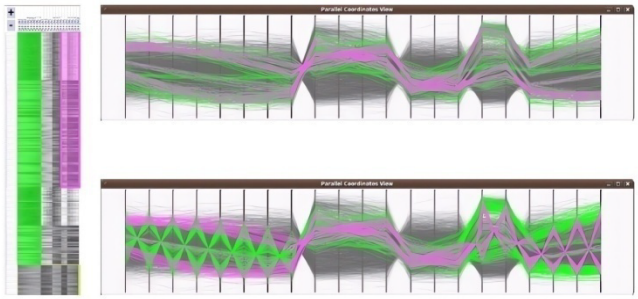


Figure 8: Two biclusters visualized as heatmaps (on the left) and next mapped to the parallel coordinates visualization without centroids (top right) then with centroids (bottom right) [51]

Parallel coordinates are a suitable method for visualizing large biclusters or individual biclusters. However, the cluttering of polylines due to overlapped biclusters can hinder the effectiveness of this technique in displaying multiple biclusters in a single view.

In general, scalability is the primary limitation of these methods (i.e., heatmaps and parallel coordinates), whether due to the abundance of biclusters or to high rates of overlap [45].

3.3. Bubble map visualization

In [45] and [52], the author presented an approach that involves the depiction of biclusters as circular entities named bubbles. The color coding of these bubbles corresponds to the sets of biclusters generated through a biclustering algorithm, with the capability to display up to three sets simultaneously. The intensity of the color, or brightness, indicates the uniformity within a bicluster. The size of each bubble is determined by the product of the number of genes and the number of conditions that constitute the bicluster. The placement of each bubble is based on a two-dimensional projection derived from multidimensional points which are the rows and columns that make up the bicluster. While this visualization method intuitively represents the arrangement of biclusters, the overlapping of bubbles does not precisely mirror the actual overlaps between biclusters. Rather, it serves as an approximation of their similarity. This technique is often employed to complement other methods, aiding in the comprehension of the general patterns observed in biclustering analyses. An illustration of this visualization method is provided in Figure 9.

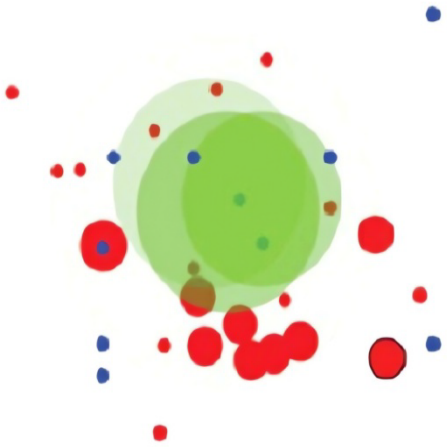


Figure 9: Bubblemap visualization of the outcomes of three biclustering algorithms [19]

Due to the limitations of heatmaps and parallel coordinates in visualizing a large number of biclusters, particularly with high levels of overlap [45], more advanced visualization approaches have been introduced. These approaches combine traditional gene expression visualization techniques (such as heatmaps and/or parallel coordinates) with set visualization techniques [53] like Venn-like diagrams [45], node-link diagrams [54] and two-dimensional matrix representations [55][56]. The following provides a description of these innovative techniques.

3.4. Venn-like diagrams visualization

Euler and Venn diagrams stand as some of the earliest techniques for illustrating sets and their interconnections. These diagrams were conceptualized by the British mathematician and philosopher, John Venn in the 18th century and have been widely adopted as effective tools for teaching concepts of set theory and logical relationships in education [57]. Utilizing a *proportional-*

area model, where the depicted areas correspond to the magnitude of a set and its intersections, sets are symbolized by enclosed *shapes* on a plane, typically *circles* and the relationships between sets are demonstrated through the *overlapping* of these shapes. They offer a versatile means to represent all conceivable set relationships, including intersection, inclusion and exclusion, due to the absence of limitations on the representation of overlaps. Venn diagrams, which are a specialized variant of Euler diagrams, are capable of representing every conceivable set intersection, regardless of whether they are non-empty or not.

In [45], the author described a novel visualization method based on Venn diagrams was employed, where biclusters are visualized as non-uniform shapes termed as *hulls* and their intersections are indicated by the *hulls' overlaps.* To represent genes and conditions that are unique to a single bicluster or shared among certain intersections, symbols known as *glyphs* are used. Each glyph is designed as a *pie chart*, segmented into sectors that represent the count of biclusters associated with the genes and conditions. The dimension of a glyph is indicative of the size of its respective group. The graphical representation is organized using a *force-directed algorithm* where biclusters are illustrated as flexible dynamic groups of genes and conditions. Specific genes and conditions related to a single bicluster or overlaps between a set of biclusters are illustrated through heatmaps and/or parallel coordinates which are displayed separately under request. While this approach effectively identifies a considerable number of biclusters with minimal rate of overlaps, its performance may degrade when faced with datasets containing a high degree of biclusters overlap. Figure 10 shows an illustration of this visualization technique.

3.5. Node-link diagrams visualization

A node-link diagram is a graphical representation, either in 2D or 3D consisting of *nodes* and connecting *edges*. This visualization technique represents various entities as nodes, also known as *vertices* and the connections between these entities as *edges* or *links*. Typically, nodes are symbolized by geometric shapes such as *circles* while the connections are depicted by *lines*. Creating a clear and informative graph requires careful consideration of nodes placement and edges routing, particularly when dealing with a large number of elements. Force-directed layout is commonly employed to address this challenge.

In [54], the author depicted biclusters and their intersections through a node-link *graph*. Here, biclusters form the nodes while the shared genes and conditions among them are represented as *edges* or *bands*. Each bicluster is visualized as a heatmap matrix where rows correspond to genes and columns represent conditions. Overlaps between biclusters are represented by connecting *bands* that link the corresponding heatmaps at the positions of shared genes and conditions. The thickness of these bands indicates the degree of overlap, with thicker bands signifying more shared elements. The graph layout employs a *force-directed algorithm* where biclusters with overlapping elements are drawn closer together. When a bicluster is selected, it reveals detailed information such as its designation or the identifiers of its genes and conditions. This visualization approach is highly interactive and straightforward since the design is based on the heatmaps visualization. The bands that indicate overlap offer the user a detailed view of the common genes and conditions found in each pair of biclusters. However, this method's visualization of overlaps on a one-to-one basis makes it challenging to distinguish multiple bicluster overlaps easily. Moreover, the technique's scalability is limited with an increase in overlap levels. In fact, the bands become overly congested, making it difficult to gain a comprehensive view of the biclustering outcomes. An example of this bicluster visualization method is presented in Figure 11.

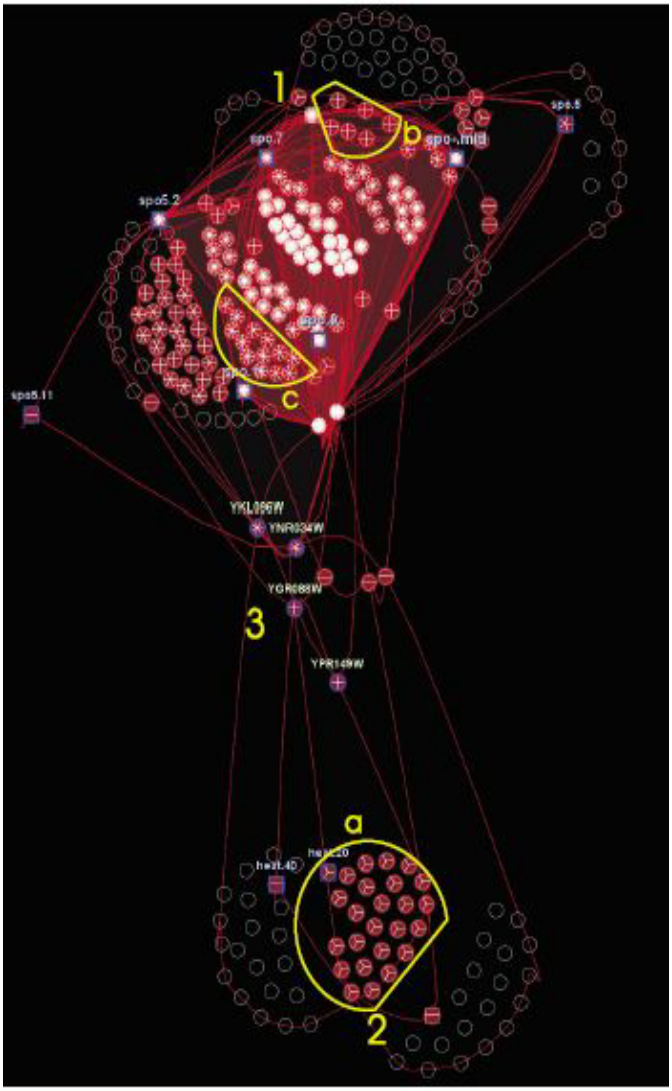


Figure 10: 50 biclusters visualization. Three sets of biclusters and their intersections are easily identifiable using the hulls representation (groups 1, 2 and 3) [45]

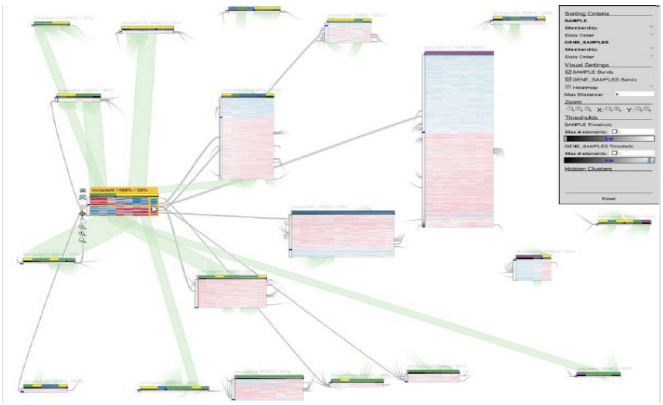


Figure 11: 20 biclusters are depicted. Nodes represent the biclusters displayed as heatmaps and edges connect the corresponding heatmaps at the positions of shared rows and columns [54]

3.6. Two-dimensional matrix visualization

In [55], [56] the author proposed a visualization method that takes into account the special characteristics of biclustering which are overlaps and bi-dimensionality. The primary objectives of the method are:

- Visually represent biclusters of varying sizes and degrees of overlap.
- Maintain both elements (i.e., genes and/or conditions) and biclusters information within a single view, preventing context loss.
- Avoid information simplification or duplication. While alternative approaches might offer clearer visualizations, they often ignore interesting information or introduce ambiguities.
- Provide interactive features that enable diverse perspectives and facilitate exploratory analysis.
- Increase scalability. An effective bicluster visualization method should accommodate large datasets, numerous biclusters and extensive overlaps between biclusters.

To achieve these objectives, the authors developed a visualization technique that lay out the generated biclusters as a *two-dimensional matrix*. Each bicluster is represented as a *column* and overlaps between sets of biclusters are depicted as *rows*. This method combines a modified set visualization technique for matrix layout with a traditional heatmap approach for visualizing individual biclusters and their overlaps as gene expression matrices [58], [53]. A user interface is implemented to query the biclusters intersection matrix and visualize matching results. The proposed technique is implemented in a web-based interactive visualization tool called *VisBicluster* which supports features like sorting, zooming and on-demand details. While applicable to any type of overlapping groups, the primary focus of this technique is on representing biclusters derived from gene expression data.

This approach emphasizes overlaps, making their identification and selection straightforward within the defined matrix-based visualization. By avoiding element crossings (lines, shapes, etc.), this method minimizes visual clutters. VisBicluster's scalability is remarkable since it can efficiently display large numbers of highly overlapped biclusters simultaneously. The tool also incorporates linking and brushing techniques for inspecting selected data subsets from different perspectives. Figure 12 illustrates this visualization technique.

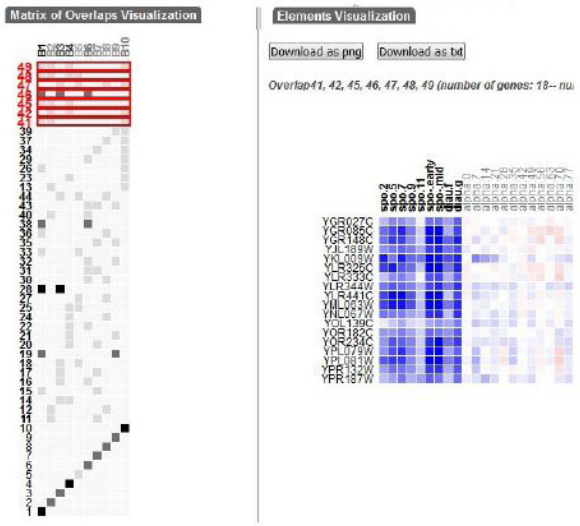


Figure 12: Plaid model biclustering algorithm [40] result visualization for yeast *Saccharomyces cerevisiae* expression data [56]. Overlaps are visualized as a two-dimensional matrix which is clustered by *Levenshtein distance*[59], an algorithm that arranges the most similar bicluster overlaps near each other (on the left). Genes and conditions of a set of similar selected overlaps are represented as a heatmap (on the right).

4. Critique of visualization methods

Our evaluation of the surveyed techniques focused on three key aspects:

- Reducing *overlaps* among biclusters.
- Maximizing the number of biclusters visualized within a single view (i.e., *scalability*).
- Ensuring *clear visibility* of both biclusters and their overlapping regions.

Due to their geometric limitations, heatmaps and parallel coordinates often struggle to efficiently visualize biclusters, particularly when evaluated against the criteria of overlap minimization and scalability [45]. Heatmaps, in particular, are typically unbalanced in terms of dimensions, with many more genes (around $10^{[3-4]}$ rows) than conditions (around $10^{[1-2]}$ columns) [19]. So, replication techniques used to visualize biclusters can lead to large matrices when visualizing multiple biclusters, making overlap perception difficult and limiting scalability. Additionally, the common use of a green-black-red color scale in heatmaps can hinder human perception of expression levels [7], [19].

Parallel coordinates often suffer from cluttering due to overlapping polylines when visualizing multiple biclusters simultaneously. This makes it difficult to perceive overlaps and limits scalability. Individual biclusters can be easily interpreted due to the human brain's ability to recognize patterns like parallel lines, mirror effects and changes in slope [19]. However, visualizing several large biclusters with high rates of overlap in the same parallel coordinates can be in some cases impossible.

By combining heatmaps and/or parallel coordinates with more sophisticated sets visualization techniques like Venn diagrams [45], node-link diagrams [54] or two-dimensional matrix visualization [55], [56], we can confirm that the scalability and clarity of drawn biclusters are improved significantly. Representing biclusters and their overlaps as abstract elements such as hulls [45], bands between heatmaps [54] or cells in a matrix [56] can simplify the visualization. So, focusing on intersections between visualized elements (i.e., biclusters) in a global overview while providing details (i.e., gene expression levels) in a separate view as heatmaps or parallel coordinates, will alleviate the representation remarkably.

Overlap is a key aspect when visualizing biclusters and overlap-centered tasks like how many overlaps between a specific number of biclusters or what are the biclusters involved in a certain overlap, can help analysts to gain insights from visualized complex data. Visualization tools that help interpret complex analysis results without distorting or losing the original data context are crucial for understanding data. In fact, VisBicluster offers an overlap-centered solution that aids in understanding biclustering results, providing scalability for analyzing real gene expression data [55], [56].

While these three novel visualization methods (i.e., BicOverlapper, Furby and VisBicluster) offer high rate of scalability, they can still be overwhelmed by a large number of biclusters and overlaps. So, in some cases, it may be impossible to visualize all biclusters within a single view. Table 2 provides a summary of biclustering visualization techniques, outlining their key characteristics.

Table 2: Evaluation of biclustering visualization techniques visualization techniques

Biclustering visualization technique	Visualization methods	Dealing with overlaps	Scalability	Clarity of visualization	Time complexity
Heatmap in [46]	Heatmap	Reordering of rows and columns	Low	Low	$O(\|X\|^2\|V\|^2) + O(\|X\|\|V\|)O(MinLA)^*$ $O(\|X\|^2\|V\|^3) + O(\|X\|\|V\|)O(MinLA)^{**}$ $O(\|X\|^2\|V\|^{2d}) + O(\|X\|\|V\|)O(MinLA)^{***}$
Heatmap in [47]	Heatmap	Reordering of rows and columns	Low	Low	$O(m^\alpha)$ with $\alpha \in [1.6, 2]^+$ $O(n^\alpha)$ with $\alpha \in [2.5, 2.7]^+$
Heatmap in [48]	Heatmap	Replication of rows and columns	Low	Low	$O(mn^2 + n^2 \log n)^\#$
Heatmap in [51]	Heatmap	Replication of rows and columns	Low	Low	–
Parallel coordinates plots [51]	Parallel coordinates plots	Color polylines	Low	Low	–
Bubblemap[45] and [52]	circles	Intersections between circles	Medium	High	–
Venn diagrams [45]	Hulls, pie charts, Heatmap, parallel coordinates	Hulls intersections, Glyphs with pie chart sectors	High	High	$O(n^3)^\$$

Node-link diagrams [54]	Heatmap, bands	Bands between rows and columns of heatmaps	High	High	–
Two-dimensional matrix visualization [55] [56]	Heatmap, two-dimensional matrix	Cells of the two-dimensional matrix	High	High	$O(n*m)^{\&}$

Where:
V is the set of vertices and *X* is the set of hyperedges.
*If we convert hyperedges into paths or cycles.
**If we convert each hyperedge into maximum valid cycles.
***If we convert each hyperedge into d arbitrary valid cycles.
*O(MinLA)*is the time complexity of the minimum linear arrangement algorithm.
^{+}m is the number of biclusters and *n* is the number of elements.
$^{\#}n$ is the number of biclusters and *m* is the number of rows and columns in all the biclusters.
$^{\$}n$ isthe number of nodes (genes and conditions).
$^{\&}n$ is the number of biclusters and m is the total number of overlaps between biclusters.

5. Tools and datasets

5.1. Tools

Several useful tools exist that incorporate many of the biclustering visualization methods discussed in this review. Table 3 provides a summary of the key features of these tools.

- BiVoc: It is a C++ implementation that incorporates two primary programs: one for the layout algorithm and another for generating the corresponding visualization image. To address the potential issue of numerous rows and columns due to replication, a user-friendly web interface allows for the selection of specific biclusters to display [48].
- ExpressionView: It is an R package that incorporates the developed ordering method and provides interactive visualization of bicluster results as heatmaps in a Flash applet format [47].
- Bicluster Viewer: It offers heatmap and parallel coordinates visualizations for representing biclusters as contiguous blocks. It also includes a range of interactive features [51].
- Biclust: It incorporates various biclustering algorithms and offers several visualization methods, including the Bubbleplot graphical representation that depicts biclusters as circles [52].
- BicOverlapper: It is a Java package that enables the visualization of bicluster sets using Venn-like diagrams, the representation of microarray data matrices or individual biclusters as heatmaps and/or parallel coordinates and the visualization of transcription regulatory networks. It also supports the integration of these various visualization techniques for comprehensive data analysis [45].
- Furby: It is a Java implementation of the node-link diagram technique for visualizing biclusters[54]. It incorporates several features including ordering, zooming, adjustable thresholds for different defined values, etc.
- VisBicluster: It is a web application built using JavaScript and the D3 library[60]. It enables the visualization of biclusters and their potential overlaps using the two-dimensional matrix method. Single biclusters or overlaps between two or more biclusters are depicted by heatmaps in a separate view. The software includes various integrated features such as ordering, filtering, zooming, etc. Linking and brushing between visualization techniques in VisBicluter are supported [55], [56].

Table 3: Characteristics of biclustering visualization tools

Tool	Heatmap	Parallel coordinates	Other visualizations	Degree of interactivity	accessibility	Available at
BiVoc [48]	Yes	No	_	Medium	free	http://bioinformatics.cs.vt.edu/~murali/papers/bivoc
ExpressionView [47]	Yes	No	_	Medium	free	http://www.unil.ch/cbg/ExpressionView

Bicluster Viewer [61]	Yes	Yes	–	High	commercial	http://www.simtech.uni-stuttgart.de
Biclust [52]	Yes	Yes	Bubbleplot, beplot, boxplot	Low	free	https://cran.r-project.org/web/packages/biclust/index.html
BicOverlapper [45]	Yes	Yes	Venn-like diagrams, TRN, word cloud	Very high	free	http://vis.usal.es/bicoverlapper
Furby [54]	Yes	No	Bar chart, histogram	Very high	free	http://furby.caleydo.org
VisBicluter [55] [56]	Yes	No	Two-diensional matrix, cells	Very high	free	http://vis.usal.es/~visusal/visbicluster

Table 4: Gene expression datasets utilized to evaluate biclustering visualization techniques

Name	Genes	Experimental conditions	Reference
Yeast*Saccharomyces cerevisiae*microarray data	2467	79	[1]
*Prostate cancer tissue*dataset	54675	19	[62]
*Human lung carcinomas*dataset	12600	203	[63]
*Multiple tissue types*dataset	5565	102	[64]

5.2. Datasets

Multiple gene expression datasets were employed to assess biclustering visualization techniques. Table 4 provides a list of some of these datasets.

6. Conclusion

While clustering focuses on identifying groups of similar elements within a dataset (gene expression data) by applying algorithms on one dimension either rows (i.e. genes) or columns (i.e. conditions), biclustering seeks to uncover patterns that exist simultaneously across both rows and columns. This added complexity (for biclustering case) necessitates more advanced visualization methods to effectively analyze gene expression data. By providing insights into the underlying relationships between genes and conditions, these visualizations can help bioinformaticians extract valuable knowledge. We present in this paper a global review that summarizes the most mentioned techniques to visualize results of biclustering of gene expression data in the literature and we evaluate them. We can mention that visualization issues such as scalability and overlaps between biclusters can be considered as open directions for academicians and researchers. As a possible solution to simplify the complexity of visualization of biclustering results, we think that the combination between traditional visualization techniques like heatmaps or parallel coordinates and one of the novel set visualization techniques mentioned in the literature [53], can be useful [56].

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, 1998, doi.org/10.1073/pnas.95.25.1486

[2] R.. Sokal, C.. Michener, "A statistical method for evaluating systematic relationships," *Univ. Kansas, Sci. Bull.*, vol. 38, , pp. 1409–1438, 1958.

[3] J.A. Hartigan, M.A. Wong, Algorithm AS 136: A K-Means

Clustering Algorithm, 1979, doi.org/10.2307/2346830

[4] Y. Cheng, G.M. Church, "Biclustering of expression data," *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 93–103, 2000.

[5] S.C. Madeira, A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004, doi:10.1109/TCBB.2004.2.

[6] B. Pontes, R. Giráldez, J.S. Aguilar-Ruiz, "Biclustering on expression data: A review," *Journal of Biomedical Informatics*, vol. 57, pp. 163–180, 2015, doi:10.1016/j.jbi.2015.06.028.

[7] C. Ware, Information visualization : perception for design, Morgan Kaufman, 2004.

[8] B.J. Fry, Computational information design, Massachusetts Institute of Technology Cambridge, MA, USA, 2004.

[9] J.J. Thomas, K.A. Cook, Illuminating the path, IEEE Computer Society, 2005.

[10] D. Keim, K. Jörn, G. Ellis, M. Florian, Mastering the information age : solving problems with visual analytics, Eurographics Association, 2010.

[11] A. Holzinger, Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What Is the Benefit of Bringing Those Two Fields to Work Together?, Springer, Berlin, Heidelberg, vol.8127, pp 319–328, 2013, doi:10.1007/978-3-642-40511-2_22.

[12] W. Ayadi, M. Elloumi, Biological Knowledge Visualization, John Wiley & Sons, Inc., Hoboken, New Jersey: 651–661, 2011.

[13] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 2, pp. 69–91, 1985, doi:10.1007/BF01898350.

[14] D. Gonçalves, R.S. Costa, R. Henriques, "Context-situated visualization of biclusters to aid decisions: going beyond subspaces with parallel coordinates," *ACM International Conference Proceeding Series*, no. 9, pp. 1–5, doi:10.1145/3531073.3531124.

[15] N.K. Verma, T. Sharma, S. Dixit, P. Agrawal, S. Sengupta, V. Singh, "BIDEAL: A Toolbox for Bicluster Analysis—Generation, Visualization and Validation," *SN Computer Science*, vol. 2, no. 1, 2021, doi:10.1007/S42979-020-00411-9.

[16] M. Sözdinler, "A Review of Visualization Methods and Tools for the Biclustering," *International Journal of Innovative Science and Research Technology*, vol. 6, 2021, doi.org/10.48550/arXiv.2111.12154.

[17] H. Aouabed, R. Santamaria, M. Elloumi, "Visualizing biclustering results on gene expression data: A survey," *ACM International Conference Proceeding Series*, pp. 170–179, 2021, doi:10.1145/3473258.3473284.

[18] H. Aouabed, M. Elloumi, R. Santamaría, "An evaluation study of biclusters visualization techniques of gene expression data," *Journal of Integrative Bioinformatics*, vol. 18, no. 4, 2021, doi:10.1515/JIB-2021-0019/MACHINEREADABLECITATION/RIS.

[19] R. Santamaria, Visual analysis of gene expression data by means of biclustering, University of Salamanca, Spain, 2009.

[20] A.V. Freitas, W. Ayadi, M. Elloumi, J. Oliveira, J. Oliveira, J.-K. Hao, Survey on Biclustering of Gene Expression Data, John Wiley & Sons, Inc., Hoboken, New Jersey: 591–608, 2012, doi:10.1002/9781118617151.ch25.

[21] H. Ben Saber, M. Elloumi, "Dna Microarray Data Analysis: a New Survey on Biclustering," *International Journal for Computational Biology*, vol. 4, no. 1, pp. 21, 2015, doi:10.34040/ijcb.4.1.2014.36.

[22] S.C. Madeira, A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[23] V.A. Padilha, R.J.G.B. Campello, "A systematic comparative evaluation of biclustering techniques," *Padilha Campello BMC Bioinforma.*, vol. 18, , 2017, doi:10.1186/s12859-017-1487-1.

[24] G. Getz, E. Levine, E. Domany, "Coupled two-way clustering analysis of gene microarray data.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, pp. 12079–84, 2000, doi:10.1073/pnas.210134797.

[25] C. Tang, L. Zhang, A. Zhang, M. Ramanathan, "Interrelated two-way clustering: an unsupervised approach for gene expression data analysis," in *Proceedings 2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2001)*, IEEE: 41–48, 2001, doi:10.1109/BIBE.2001.974410.

[26] S. Busygin, G. Jacobsen, E. Krämer, "Double Conjugated Clustering Applied to Leukemia Microarray Data," *IN 2ND SIAM ICDM, WORKSHOP ON CLUSTERING HIGH DIMENSIONAL DATA*, 2002.

[27] J.A. Hartigan, "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123, 1972, doi:10.2307/2284710.

[28] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006, doi:10.1093/bioinformatics/btl060.

[29] A. Ben-Dor, B. Chor, R. Karp, Z. Yakhini, "Discovering local structure in gene expression data," *Proceedings of the Sixth Annual International Conference on Computational Biology - RECOMB '02*, pp. 49–57, 2002, doi:10.1145/565196.565203.

[30] T.M. Murali, S. Kasif, "Extracting conserved gene expression motifs from gene expression data.," *Pacific Symposium on Biocomputing.*, vol. 88, , pp. 77–88, 2003, doi:10.1142/9789812776303_0008.

[31] S. Bergmann, J. Ihmels, N. Barkai, "Iterative signature algorithm for the analysis of large-scale gene expression data," *Physical Review E*, vol. 67, no. 3 1, pp. 031902/1-031902/18, 2003, doi:10.1103/PhysRevE.67.031902.

[32] H. Cho, I.S. Dhillon, Y. Guan, S. Sra, "Minimum Sum-Squared Residue Co-clustering of Gene Expression Data," in *Proceedings of the 2004 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Philadelphia, PA: 114–125, 2004, doi:10.1137/1.9781611972740.11.

[33] G. Li, Q. Ma, H. Tang, A.H. Paterson, Y. Xu, "QUBIC: A qualitative biclustering algorithm for analyses of gene expression data," *Nucleic Acids Research*, vol. 37, no. 15, 2009, doi:10.1093/nar/gkp491.

[34] C. Huttenhower, K. Tsheko Mutungu, N. Indik, W. Yang, M. Schroeder, J.J. Forman, O.G. Troyanskaya, H.A. Coller, "Detailing regulatory networks through large scale data integration," *Bioinformatics*, vol. 25, no. 24, pp. 3267–3274, 2009, doi:10.1093/bioinformatics/btp588.

[35] D. Bozdag, J.D. Parvin, U. V Catalyurek, "A biclustering method to discover co-regulated genes using diverse gene expression datasets," *International Conference on Bioinformatics and Computational Biology*, vol. 5462 LNBI, , pp. 151–163, 2009, doi:10.1007/978-3-642-00727-9_16.

[36] A.A. Shabalin, V.J. Weigman, C.M. Perou, A.B. Nobel, "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 985–1012, 2009, doi:10.1214/09-AOAS239.

[37] A. Tanay, R. Sharan, R. Shamir, "Discovering statistically significant biclusters in gene expression data.," *Bioinformatics*, vol. 18 Suppl 1, , pp. S136-S144, 2002, doi:10.1093/bioinformatics/18.suppl_1.S136.

[38] D.S. Rodriguez-Baena, A.J. Perez-Pulido, J.S. Aguilar-Ruiz, "A biclustering algorithm for extracting bit-patterns from binary datasets," *Bioinformatics*, vol. 27, no. 19, pp. 2738–2745, 2011, doi:10.1093/bioinformatics/btr464.

[39] A. Serin, M. Vingron, "DeBi: Discovering Differentially Expressed Biclusters using a Frequent Itemset Approach.," *Algorithms Molucular Biology*, vol. 6, no. 1, pp. 18, 2011, doi:10.1186/1748-7188-6-18.

[40] L. Lazzeroni, A. Owen, "Plaid Models for Gene Expression Data," *CEUR Workshop Proc.*, vol. 1542, , pp. 33–36, 2000, doi:10.1017/CBO9781107415324.004.

[41] Y. Kluger, R. Basri, J.T. Chang, M. Gerstein, "Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions," *Genome Research*, vol. 13, pp. 703–716, 2003, doi:10.1101/gr.648603.graph.

[42] J. Gu, J.S. Liu, "Bayesian biclustering of gene expression data," *BMC Genomics*, vol. 9 Suppl 1, pp. S4, 2008, doi:10.1186/1471-2164-9-S1-S4.

[43] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. van Sanden, D. Lin, W. Talloen, L. Bijnens, H.W.H. Göhlmann, Z. Shkedy, D.A. Clevert, "FABIA: Factor analysis for bicluster acquisition," *Bioinformatics*, vol. 26, no. 12, pp. 1520–1527, 2010, doi:10.1093/bioinformatics/btq227.

[44] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, E. Zitzler, "BicAT: A biclustering analysis toolbox," *Bioinformatics*, vol. 22, no. 10, pp. 1282–1283, 2006, doi:10.1093/bioinformatics/btl099.

[45] R. Santamaría, R. Therón, L. Quintales, "A visual analytics approach for understanding biclustering results from microarray data," *BMC Bioinformatics*, vol. 9, no. 1, pp. 247, 2008, doi:10.1186/1471-2105-9-247.

[46] R. Jin, Y. Xiang, D. Fuhry, F.F. Dragan, "Overlapping Matrix Pattern Visualization: A Hypergraph Approach," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE: 313–322, 2008, doi:10.1109/ICDM.2008.102.

[47] A. Luscher, G. Csardi, A. Morton de Lachapelle, Z. Kutalik, B. Peter, S. Bergmann, "ExpressionView--an interactive viewer for modules identified in gene expression data," *Bioinformatics*, vol. 26, no. 16, pp. 2062–2063, 2010, doi:10.1093/bioinformatics/btq334.

[48] G.A. Grothaus, A. Mufti, T. Murali, "Automatic layout and visualization of biclusters," *Algorithms for Molecular Biology*, vol. 1, no. 1, pp. 15, 2006, doi:10.1186/1748-7188-1-15.

[49] S. Batzoglou, S. Istrail, Physical Mapping with Repeated Probes: The Hypergraph Superstring Problem, Springer, Berlin, Heidelberg: 66–77, 1999, doi:10.1007/3-540-48452-3_5.

[50] K.S. Booth, G.S. Lueker, "Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms," *Journal of Computer and System Sciences*, vol. 13, no. 3, pp. 335–379, 1976, doi:10.1016/S0022-0000(76)80045-1.

[51] J. Heinrich, R. Seifert, M. Burch, D. Weiskopf, BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data, Springer, Berlin, Heidelberg: 641–652, 2011, doi:10.1007/978-3-642-24028-7_59.

[52] S. Kaiser, R. Santamaria, T. Khamiakova, M. Sill, R. Theron, L. Quintales, F. Leisch, E. De, T. Maintainer, "biclust: BiCluster Algorithms. R package version 1.0.2," 2013.

[53] H. Aouabed, R. Santamaría, M. Elloumi, Suitable Overlapping Set Visualization Techniques and Their Application to Visualize Biclustering Results on Gene Expression Data, Springer, Cham: 191–201, 2018, doi:10.1007/978-3-319-99133-7_16.

[54] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, S. Hochreiter, "Furby: fuzzy force-directed bicluster visualization.," *BMC Bioinformatics*, vol. 15 Suppl 6, no. Suppl 6, pp. S4, 2014, doi:10.1186/1471-2105-15-S6-S4.

[55] H. Aouabed, R. Santamaria, M. Elloumi, "VisBicluster: A Matrix-Based Bicluster Visualization of Expression Data," *Journal of Computational Biology*, pp. cmb.2019.0385, 2020, doi:10.1089/cmb.2019.0385.

[56] H. Aouabed, M. Elloumi, "Visualizing Biclusters of Gene Expression Data and Their Overlaps Based on a Two-Dimensional Matrix Technique," *Computational Biology and Bioinformatics 2023*, vol. 11, no. 2, pp. 19–32, 2023, doi:10.11648/J.CBB.20231102.11.

[57] M.E. Baron, "A Note on the Historical Development of Logic Diagrams: Leibniz, Euler and Venn," *The Mathematical Gazette*, vol. 53, no. 384, pp. 113, 1969, doi:10.2307/3614533.

[58] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, "UpSet: Visualization of intersecting sets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1983–1992, 2014, doi:10.1109/TVCG.2014.2346248.

[59] V.I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Sov. Phys. Dokl. Vol. 10, p.707*, vol. 10, , pp. 707, 1966.

[60] M. Bostock, V. Ogievetsky, J. Heer, "D^3 Data-Driven Documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011, doi:10.1109/TVCG.2011.185.

[61] J. Heinrich, M. Burch, R. Seifert, D. Weiskopf, "BiCluster Viewer : A Visualization Tool for Analyzing Gene Expression Data BiCluster Viewer : A Visualization Tool for Analyzing Gene Expression Data," 2011.

[62] S. Varambally, J. Yu, B. Laxman, D.R. Rhodes, R. Mehra, S.A. Tomlins, R.B. Shah, U. Chandran, F.A. Monzon, M.J. Becich, J.T. Wei, K.J. Pienta, D. Ghosh, M.A. Rubin, A.M. Chinnaiyan, "Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression," *Cancer Cell*, vol. 8, no. 5, pp. 393–406, 2005, doi:10.1016/j.ccr.2005.10.001.

[63] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Acad Sciences U. S. A.*, vol. 98, no. 24, 13790–13795, 2001, doi:10.1073/pnas.191502998.

[64] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, J.B. Hogenesch, "Large-scale analysis of the human and mouse transcriptomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 7, 4465–4470, 2002, doi:10.1073/pnas.012025199.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



HAITHEM AOUABED received an Undergraduate Degree in Mathematics in 2006, and a Master's Degree in Computer Science in 2010, from the Faculty of Economic Sciences and Management, Sfax, Tunisia. He also received a Master's Degree in Computer Science in 2012 and a PhD Degree in Computer Science in 2024, from the University of Sfax, Tunisia. His research interests include Bioinformatics and Information Visualization, especially the integration of different source

data, analysis algorithms and representations for a better understanding of biological problems.



MOURAD ELLOUMI received an Undergraduate Degree in Mathematics and Physics in 1984, and a master's degree in computer engineering in 1988, from the Faculty of Sciences of Tunis, Tunisia. He also received a Master's Degree in Computer Science in 1989 and a PhD Degree in Computer Science in 1994, from the University of Aix-Marseilles III, France. Then, he received a *Habilitation* for conducting research in Computer Science in 2003, from the National School of Computer Science, Tunis, Tunisia. He is currently a Full Professor in Computer Science, College of Computing and Information Technology, University of Bisha, Saudi Arabia. Professor Mourad Elloumi is the author/co-author of more than 80 publications in international journals, books and conference proceedings. He was a Guest Editor of a special issue on biological knowledge discovery and data mining, *Knowledge Based Systems* Journal (Elsevier 2002), a Guest Editor of a special issue on pattern finding in Computational Molecular Biology, *Recent Patents on DNA and Gene Sequence* Journal (Bentham Science 2012), a Co-Editor of the proceedings of two international conferences and Editor/Co-Editor of five books, respectively, on Algorithms in Computational Molecular Biology (Wiley 2011), Biological Knowledge Discovery (Wiley 2014), Pattern Recognition in Computational Molecular Biology (Wiley 2015), Algorithms for Next-Generation Sequencing Data (Springer 2017), Deep Learning for Biomedical Data Analysis (Springer 2021). His research interests include Algorithmics, Computational Molecular Biology, Knowledge Discovery and Data Mining, and Deep Learning.



FAHAD ALGARNI received a bachelor's degree (Hons.) from the Department of Computer Science, King Abdulaziz University, the M.I.T. degree in computer networks from La Trobe University, Melbourne, Australia, and the Ph.D. degree from the Clayton School of Information Technology, Monash University, Melbourne, Australia. He is currently the Dean of the College of Computing and Information Technology, University of Bisha, Saudi Arabia. His research interests include Wireless Sensor Networks; Cloud Computing, Systems, Design, and Reliability, the IoT, IoV, IoD, Cybersecurity and Bioinformatics.

Received: 24 September, 2024, Revised: 22 November, 2024, Accepted: 23 November, 2024, Online: 18 December, 2024

DOI: https://doi.org/10.55708/js0312003

Advanced Cloud-Based Solutions for Peripheral Artery Disease: Diagnosis, Analysis, and Visualization

Mohammed A. AboArab[1,2], Vassiliki T. Potsika[1], Dimitrios I. Fotiadis*[1,3]

[1]Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, GR45110, Greece

[2]Electronics and Electrical Communication Engineering Dept., Faculty of Engineering, Tanta University, Tanta, Egypt

[3]Biomedical Research Institute, Foundation for Research and Technology-Hellas, University Campus of Ioannina, Ioannina, GR45110, Greece

E-mail: m.aboarab@uoi.gr / vpotsika@uoi.gr / fotiadis@uoi.gr

*Corresponding author: Dimitrios I. Fotiadis, Ioannina, GR45110, Greece, +302651005580 & fotiadis@uoi.gr

ABSTRACT: Peripheral artery disease (PAD) affects 237 million people globally, leading to significant morbidity and mortality. Traditional diagnostic methods are invasive, costly, and require specialized expertise, emphasizing the need for more accessible, and accurate alternatives. This paper introduces the DECODE cloud platform, an advanced tool that leverages cloud computing, machine learning, and high-performance data visualization to enhance PAD diagnosis and treatment. The platform integrates modules for peripheral artery segmentation, reconstruction, and comprehensive data warehousing, supporting 2D and three-dimensional (3D) rendering visualization. It enables the simulation and optimization of drug-coated balloons, enhancing clinical decision-making through robust data analytics. The evaluation metrics demonstrate the platform's efficacy: the multiplanar visualization module achieved a performance score 94%, and the 3D rendering module scored 89%, with both modules attaining perfect scores in best practices and search engine optimization. These results highlight the DECODE platform's capacity to provide scalable, noninvasive diagnostic solutions, setting a new standard in digital health technologies for PAD. This study underscores the transformative potential of integrating advanced visualization and computing techniques in medical diagnostics.

KEYWORDS: Digital health technologies, peripheral artery disease, cloud computing, medical imaging visualization, noninvasive diagnostics

1. Introduction

Peripheral artery disease (PAD) is a globally prevalent circulatory disorder characterized by the narrowing of peripheral arteries, which leads to reduced blood flow to the limbs [1]. This reduction in blood flow can cause symptoms such as leg pain or claudication during physical activity, and slow or non-healing wounds on the feet or toes. In more advanced stages, PAD can result in critical limb ischemia, leading to severe pain at rest, and in some cases, the need for amputation. This condition significantly impacts mobility and quality of life and poses serious health risks. Worldwide, PAD is estimated to affect approximately 7.4% of individuals in high-income countries and 5.1% of those in low-income countries. In Europe, particularly in high-income regions, the prevalence is higher than 10% among people aged 65 and older, emphasizing the disease's burden on elderly populations [2]. In the United States, PAD affects 8–10 million individuals, with projections suggesting that this number could rise to 19 million by 2050 [3]. The global PAD market, valued at USD 4.45 billion in 2022, is expected to reach USD 9.15 billion by 2031, growing at a compound annual growth rate (CAGR) 8.35% [4, 5]. This economic growth underscores the increasing demand for effective diagnostic and therapeutic solutions for PAD. Early and accurate diagnosis of PAD is essential to mitigate its high morbidity and mortality rates.

Traditional diagnostic methods, such as digital subtraction angiography (DSA) and magnetic resonance angiography (MRA), provide high-resolution imaging of arterial structures and are considered the gold standard for PAD diagnosis. However, DSA is invasive and

involves catheterization and contrast injection, which poses risks such as bleeding, infection, and radiation exposure. While MRA is noninvasive and avoids ionizing radiation, it is expensive. Both methods are resource intensive, rely on advanced infrastructure and expertise, underscoring the need for cost-effective, scalable, and noninvasive diagnostic alternatives [6].

Recent advancements in cloud computing and machine learning are revolutionizing healthcare by offering innovative solutions to the challenges posed by traditional diagnostic methods for PAD [7, 8]. Cloud-based platforms are transforming the way health data are managed, enabling secure storage, rapid processing, and sophisticated analysis of vast datasets [9]. These platforms facilitate real-time diagnostic insights, which are crucial for timely and accurate medical decision-making. Moreover, the integration of advanced visualization techniques, such as multiplanar and three-dimensional (3D) rendering, within cloud platforms enhances the ability to examine complex anatomical structures in detail. This synergy between cloud technology and machine learning in medicine improves diagnostic accuracy and supports personalized patient care by allowing the continuous and scalable application of advanced analytics and visualization across diverse clinical settings.

In this work, the DECODE cloud platform is presented, which addresses the complexities of PAD diagnosis and treatment by integrating cutting-edge data processing, analysis, and visualization capabilities. The platform supports the simulation and optimization of drug-coated balloons (DCBs), which are increasingly used in PAD treatment. Through robust data analytics, the DECODE platform enhances clinical decision-making, allowing users to upload, process, and analyze extensive datasets, perform image reconstruction, and utilize computational modeling and machine learning techniques to accurately predict disease progression, assess treatment outcomes, and tailor personalized therapeutic strategies for PAD patients. The platform's ability to integrate these advanced technologies positions it as a pivotal tool in improving the diagnosis and management of PAD, aligning with the projected growth and evolving needs of the global PAD market.

2. Related Work

The *YORwalK* app was developed to promote exercise and track walking ability in PAD patients; however, its effectiveness requires further validation through direct patient feedback and clinical trials [10]. *TrackPAD*, a mobile intervention for supervised exercise therapy (SET), demonstrated promising initial results; however, it faced limitations due to a small sample size and short follow-up period, necessitating larger, long-term studies [11]. A smartphone-enabled exercise program introduced in the smart step trial encountered challenges in maintaining participant engagement and measuring adherence, particularly in low-resource settings [12]. Another study highlighted the need for better disease literacy and SET support in mobile interventions; however, it was constrained by a small sample size [13]. A mobile app designed for supervised home-based exercise therapy aimed to facilitate remote interventions; however, it involved minimal patient interaction, compromising data quality and requiring further development for effective supervision [14]. Research exploring machine learning (ML) and artificial intelligence (AI) applications to improve PAD outcomes emphasized critical issues such as data interoperability, algorithm bias, and the need for extensive validation [15]. An intelligent oscillometric system for PAD detection achieved high accuracy; however, it did not account for the effects of age, necessitating further validation across diverse age groups [16]. A smartphone app compared with motivational interviewing (MI) for increasing walking distance and weight loss in PAD patients found MI to be more effective; however, the study faced limitations, including a small sample size, app usage variability, and lack of an iOS version [17]. The *HOBBIT-PAD* platform was introduced to promote exercise among PAD patients; however, it is currently available only on Android and requires further evaluation by healthcare providers [18]. Moreover, a comprehensive review of SET's role in managing PAD emphasized the potential of digital health technologies for home-based interventions. However, despite their promise, these technologies suffer from low participation rates and require optimization and evaluation of long-term outcomes [19].

These studies underscore the potential of digital health technologies for PAD management but emphasize the need for larger, more comprehensive studies and further validation. Our research seeks to address these limitations by developing the DECODE cloud platform, which includes multiple modules: a PAD segmentation and reconstruction module, a data warehouse module, and a multiplanner reconstruction module featuring 2D and 3D WebGL visualization. On the basis of RESTful APIs [20], the DECODE cloud platform enables scalability and modular integration, offering a comprehensive approach to enhance diagnosis, data management, and visualization in PAD treatment, ultimately improving patient outcomes with robust and scalable solutions.

3. Materials and Methods

3.1. Data Provider

The DECODE cloud platform relies on a data provider to ensure the availability and accessibility of medical data essential for simulations, optimizations, and treatment planning [21].

Figure 1, the data provider functions as a critical bridge between the platform's core modules and various analytical tools. It integrates data from diverse sources, including medical databases, patient records, imaging systems, and research studies, harmonizing these into a unified repository for analysis and visualization.

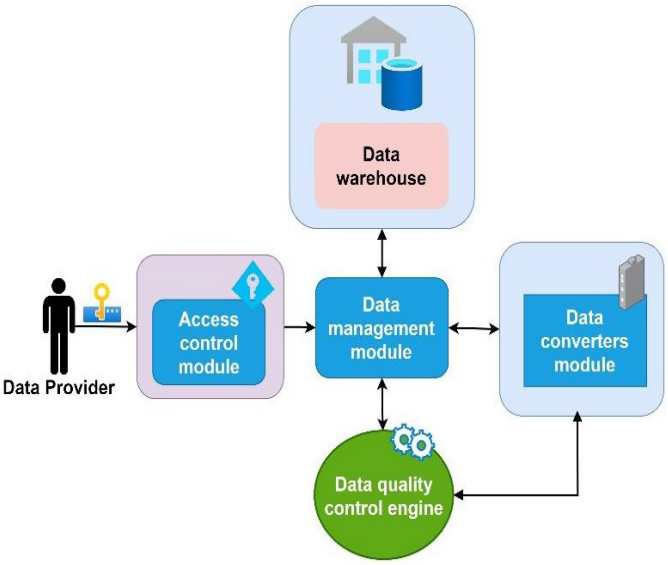


Figure 1: Data Provider Architecture within the DECODE Cloud Platform.

To maintain accuracy, the data provider continuously updates the platform with real-time data, enabling informed decision-making on the basis of the latest medical insights. The process includes cleaning, normalizing, and structuring raw medical data to ensure consistency and reliability. Quality control measures are implemented to identify and rectify any inconsistencies, enhancing the platform's reliability [22].

Security and privacy are prioritized through strict authentication and authorization protocols, allowing only authorized users to access sensitive data. The data provider supports advanced analytics, enabling the extraction of insights from complex datasets, which informs clinical decisions. As the platform scales, the data provider ensures optimal performance through efficient data retrieval and caching strategies, maintaining responsiveness even with large datasets.

3.2. *DECODE Cloud Platform Layers*

The DECODE cloud platform is a transformative solution designed to revolutionize the treatment of PAD via DCBs. By enhancing DCB simulation and optimization, the platform integrates advanced technologies and innovative research methodologies to improve treatment outcomes and patient care [23]. The conceptual multilayer hierarchical framework of the DECODE cloud platform, which provides a comprehensive overview of its architecture, is illustrated in Figure 2.



Figure 2: The DECODE Cloud Platform's Multi-Layered Hierarchical Framework.

3.2.1. *Front-End Layer: User-Centric Interactions*

The front–end layer provides clinicians with intuitive, user-friendly tools for interacting with complex medical data. These components enable seamless access to advanced visualization and analysis functionalities, enhancing clinical decision-making and improving the overall user experience.

3.2.2. *Back-End Layer: Computational Core*

The back-end layer forms the computational backbone of the platform. It incorporates advanced machine learning algorithms for automated diagnostics, computational modeling tools for simulating and optimizing PAD-specific interventions, and scalable data management modules for handling large, multi-modal datasets. Delivered as SaaS, this layer ensures high-performance processing and accurate, data-driven therapeutic strategies for PAD management.

3.2.3. *Workflow Layer: Seamless Integration and Management*

The workflow layer acts as a central hub for coordinating platform operations. Functioning as Platform-as-a-Service (PaaS), it includes a workflow management module to streamline data processing, a data quality control engine to ensure dataset accuracy, and a Docker engine and REST API manager [25] to facilitate efficient containerized application management and secure API interactions. This layer ensures seamless integration and operational consistency across the platform's processes.

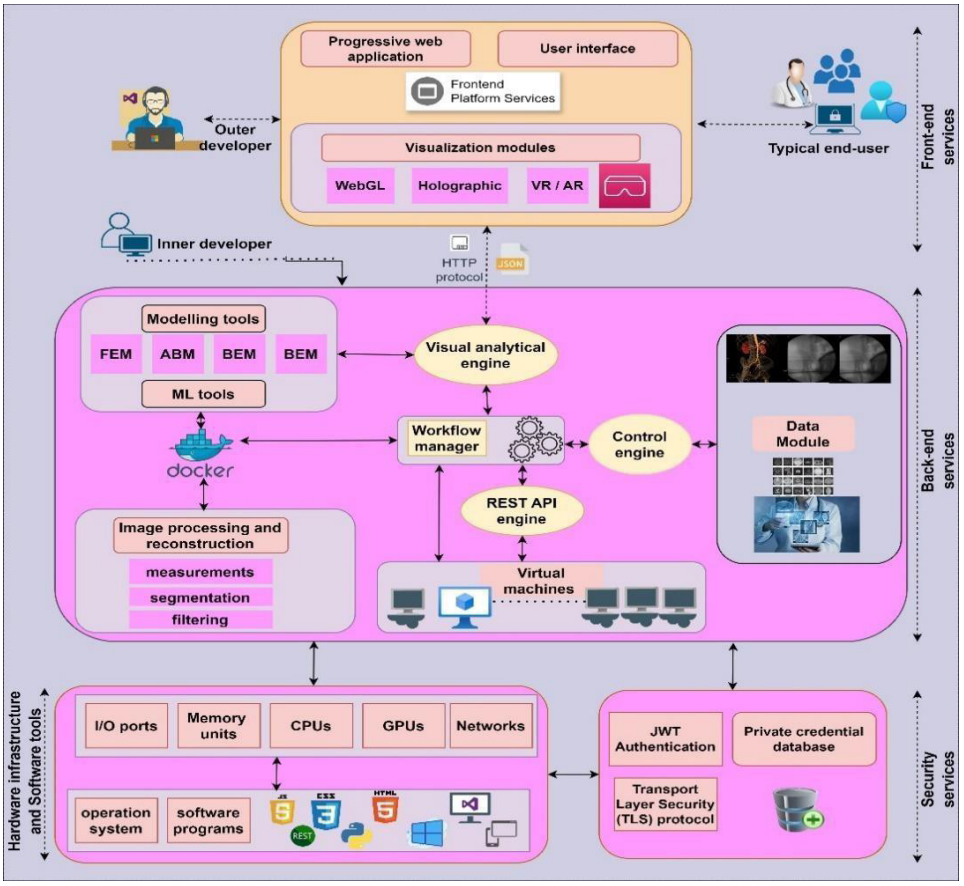


Figure 3: Comprehensive Architecture and Service Distribution of the DECODE Cloud Platform.

3.2.4. *Security Layer: Robust Data Protection*

The security layer implements advanced data protection measures, including secure socket layer (SSL) and transport layer security (TLS) protocols for encrypting data in transit, ensuring that sensitive patient information remains confidential during transfer between the platform's components. In addition, the layer employs the JSON Web Token (JWT) API framework for robust user authentication and authorization, allowing granular access control on the basis of user roles and ensuring that only authorized personnel can access specific data and functionalities within the platform.

3.2.5. *Hardware Layer: High-Performance Infrastructure*

The hardware layer, categorized as infrastructure as a service (IaaS), consists of high-performance components designed for optimal efficiency and reliability. It includes multicore dedicated servers for parallel processing, and the storage infrastructure combines SSDs for rapid data access and HDDs for cost-effective long-term storage, while GPUs are employed to expedite computationally intensive tasks such as image rendering and machine learning model training. By classifying the DECODE cloud platform's layers according to cloud computing service models, this study elucidates how the platform leverages cloud technology to provide scalable, accessible, and secure services for stakeholders, thereby opening new possibilities in PAD treatment and fostering collaboration among healthcare professionals.

3.3. *Distribution of the DECODE Cloud Computing Services to Stakeholders*

The DECODE cloud platform's architecture, as illustrated in Figure 3, is strategically designed to distribute its computing services efficiently across a diverse range of stakeholders, including clinicians, researchers, and developers. This distribution model ensures that each stakeholder can fully utilize the platform's capabilities to achieve their specific objectives in the management and treatment of PAD.

At the forefront of this distribution strategy are end-user services, which are tailored to meet the needs of clinicians and medical professionals. These services are delivered through user interfaces and visual representation modules, such as WebGL for 3D rendering, multiplanar visualization (MPV), holographic, and VR/AR technologies. These tools enable users to interact directly with patient data, provide detailed analyses and facilitate informed decision-making in real time. Front-end services, accessible via progressive web applications (PWAs) [26], form the primary touchpoint for end-users, offering an intuitive and seamless experience.

Research and development services are critical aspects of the platform's distribution strategy. These services support researchers and developers by providing access to

advanced modeling tools, ML algorithms, and robust data processing capabilities. These tools are integrated within backend services, allowing researchers to conduct complex simulations, analyze clinical data, and develop new treatment methodologies.

The workflow and integration services within the DECODE platform are designed to increase operational efficiency and promote seamless collaboration among internal teams [25]. These services, which are managed by components such as the workflow manager and the REST API engine, orchestrate the flow of tasks and data across the platform modules, ensuring seamless integration of the components. By automating critical processes such as data quality control and enabling smooth integration between various analytical tools, the platform ensures a streamlined approach to patient care.

Finally, the infrastructure services layer provides the necessary computational resources that underpin the platform's operation [27]. This layer includes essential hardware components such as I/O ports, memory units, CPUs, GPUs, and networking equipment, as well as software components and operating systems. These resources are crucial for handling the platform's computational demands, ensuring that it remains responsive and capable of scaling to meet the growing needs of its users.

4. Visual Representation Modules

The implementation architectures for the MPV and 3D rendering modules within the DECODE cloud platform are illustrated in Figure 4. These modules are integral for providing detailed and interactive visualization of medical imaging data, which are specifically tailored for diagnosing and treating PAD.

4.1. Multiplanar Visualization

4.1.1. Design and Implementation

The MPV module enables the reconstruction and visualization of 3D anatomical structures from 2D image slices, which are typically obtained from modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) [28]. The implementation of the MPV module is centered around PWAs, ensuring seamless access and interaction across various devices and platforms. The MPV module is implemented via modern web technologies, including React [29] and Cornerstone.js [30], ensuring a responsive and user-friendly interface.

These services are accessible via a PWA, which ensures seamless functionality across various devices without requiring installation. A service worker is employed to manage background tasks, such as caching and retrieving data. This enhances the performance and reliability of the PWA, allowing for offline access and faster loading times [26]. Cached data are stored in Cache Storage, ensuring that frequently accessed information is readily available.

The module uses IndexDB for local storage, allowing the browser to store large amounts of structured data [31]. This is essential for managing the vast datasets involved in medical imaging. The data are retrieved via GET requests and processed in JSON format, ensuring compatibility and ease of manipulation. The frontend platform services interact with the visualization components to render multiplanar views of the DICOM images. Users can navigate through different planes (axial,

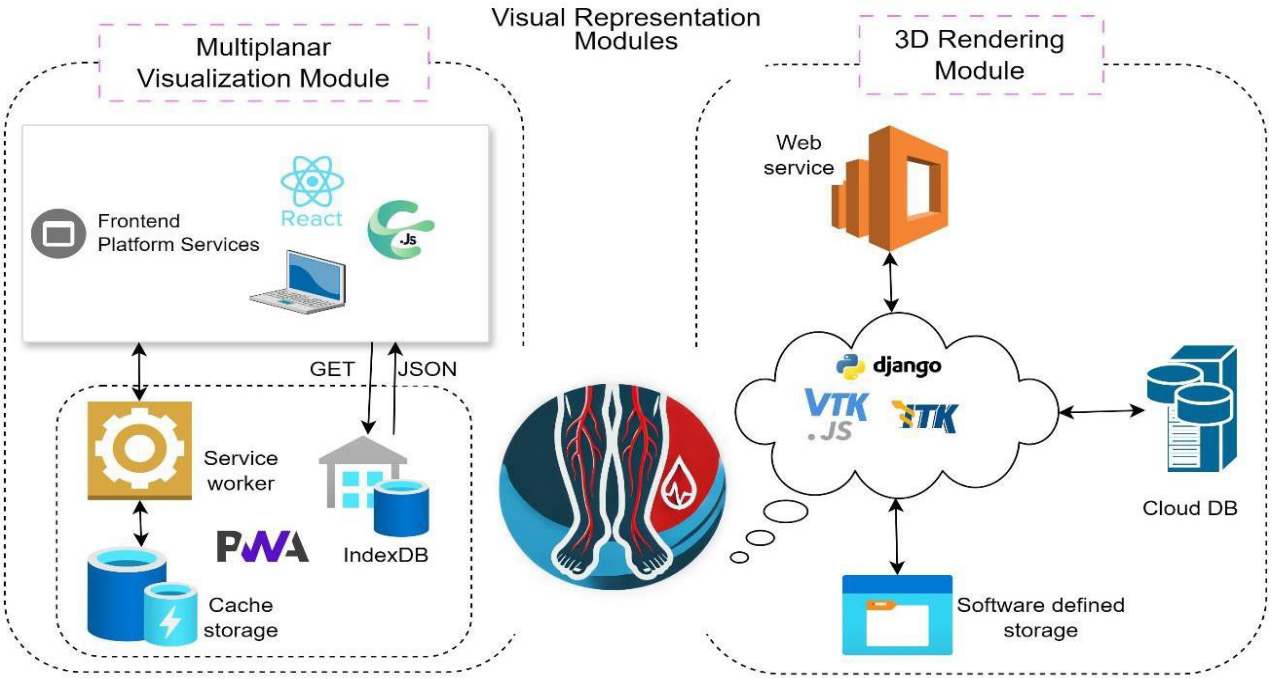


Figure 4: Implementation Architecture for Multiplanar Visualization and 3D Rendering Modules within the DECODE Cloud Platform.

sagittal, and coronal) to examine specific regions of interest. This functionality is crucial for accurate diagnosis and treatment planning.

The MPV module is built via a robust architectural framework designed to facilitate a smooth transition from data upload to advanced visualization. The key components of this framework include metadata and image information retrieval, DICOM tool activation, and DICOM image loading functionalities, which are crucial for efficient user experience and functionality.

Enhanced DICOM tools, including tools for magnification, angle measurement, and windowing, are integrated into the MPV module to complement the PWA paradigm. These tools leverage the capabilities of the Cornerstone library, which provides expanded functionalities and compatibility with various image sources. DICOM image loading is managed through the Cornerstone setup, incorporating robust success and failure handling mechanisms.

4.1.2. *User Interaction and Visualization*

The PWA approach enables seamless DICOM store connectivity, allowing users to search for and load specific studies for detailed analysis. Interactive viewers control support zooming, panning, and resetting, improving user engagement and interaction. The functionality of reference lines and planes is tailored to create spatial relationships within the MPR views, facilitating the alignment and comparison of DICOM images. This integration of reference lines involves constructing 3D lines and planes and coordinating transformations to convert 3D patient points to 2D pixel coordinates [32].

By leveraging these advanced technologies and methodologies, the MPV module within the DECODE platform offers a comprehensive and robust solution for the visualization and analysis of complex medical data.

4.1.3. *3D Rendering Module*

The 3D rendering module of the DECODE cloud platform is designed to provide high-fidelity, interactive visualizations of complex medical imaging data. This module leverages advanced WebGL technology to transform 2D CT image slices into detailed 3D representations, enhancing the ability of clinicians to assess and plan treatments with precision [33].

4.1.4. *Design and Implementation*

The 3D rendering module is crucial for enhancing the diagnostic capabilities for PAD by providing a comprehensive view of the anatomical structures. The web service, represented by a cloud-based architecture, handles the computational demands of rendering and data processing. This setup ensures scalability and efficient resource utilization, which are critical for handling large medical datasets.

The backend services of the 3D rendering module are powered by the Django framework [34], which is known for its robustness in managing heavy computational tasks. Django coordinates the processing and storage of medical imaging data, ensuring seamless integration with the front-end components. This integration is essential for providing a responsive and interactive user experience.

The core rendering functionalities are implemented using vtk.js [35] and itk.js [36], JavaScript libraries specifically designed for medical image processing and visualization. These toolkits provide extensive capabilities for 3D graphics, enabling the creation of interactive and detailed visualizations. The use of these libraries ensures that the rendered images are accurate and high-quality, which is crucial for effective diagnosis and treatment planning.

The processed data are stored in a cloud database, ensuring secure and scalable data management. In addition, software-defined storage solutions are utilized to handle dynamic storage requirements, providing flexibility and efficiency. This approach ensures that the platform can manage large volumes of data efficiently without compromising performance or security.

The 3D rendering module follows a systematic workflow, beginning with the input of volumetric data composed of multiple CT slices. These data undergo a hardware estimation process to ensure the optimal allocation of computational resources, promoting efficient data handling and processing.

4.1.5. *User Interface and Controls*

The module's user interface is designed for seamless interaction, offering controls for panning, zooming, and rotating the 3D models. These controls are essential for exploring volumetric data from various angles and perspectives, ensuring that clinicians can thoroughly inspect anatomical structures. In addition, interactive widgets enable users to adjust visualization parameters such as color mapping and opacity, allowing customized views that highlight different tissues and structures as needed [37].

By integrating these advanced technologies and methodologies, the 3D rendering module within the DECODE cloud platform significantly enhances the visualization and analysis of peripheral artery CT images. This module supports clinicians in making informed decisions, ultimately improving patient outcomes and advancing the use of CT and angiography in the diagnosis and treatment of PAD.

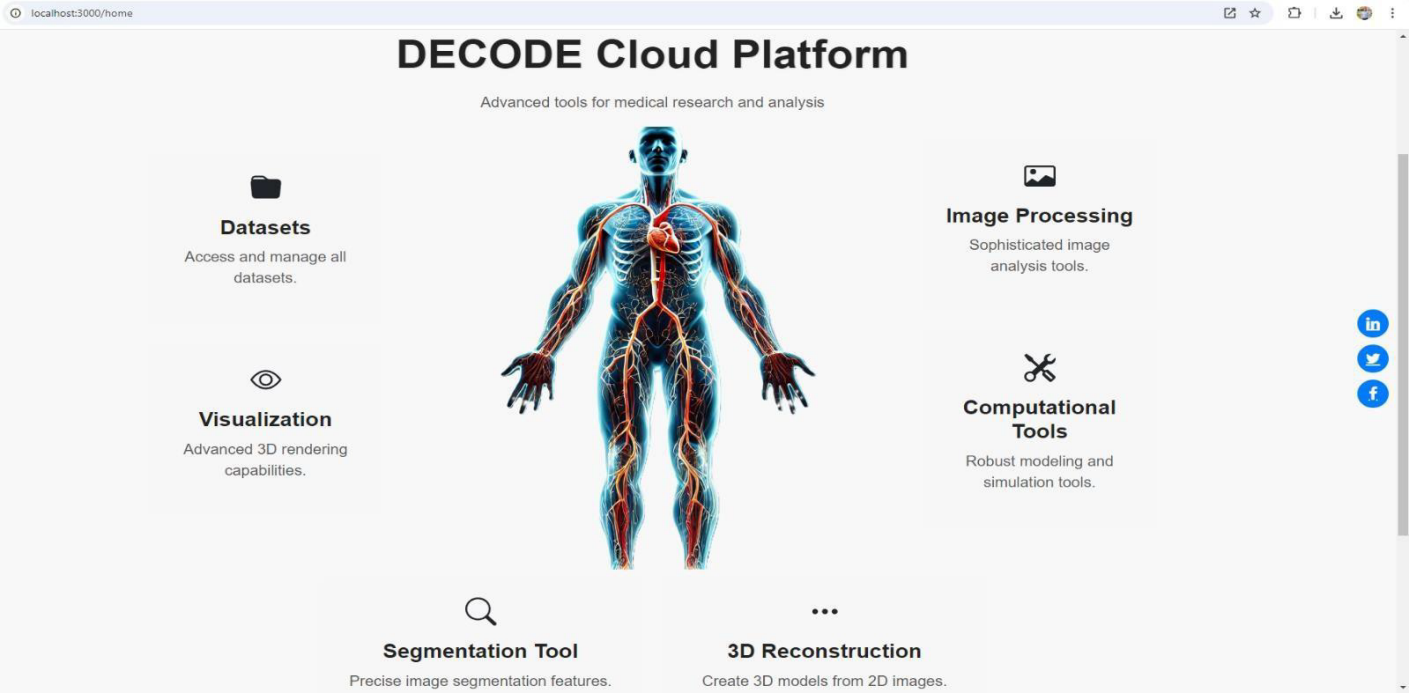


Figure 5: DECODE Cloud Platform Home Interface.

4.2. Evaluation Metrics

To assess the effectiveness and quality of the visual representation modules of the DECODE cloud platform, an evaluation was conducted via Google's Lighthouse tool [38]. Lighthouse is an open-source tool designed to evaluate the performance, accessibility, and best practices of web applications. These metrics were selected to ensure that the platform meets the critical demands of healthcare applications, which require efficiency, inclusivity, and adherence to industry standards.

The evaluation focused on two core modules of the DECODE platform: the MPV module and the 3D rendering module. Each module was tested to determine its performance in several key areas:

- Performance: This metric evaluates the speed and responsiveness of the platform, which is crucial for real-time medical imaging applications where delays can impact clinical decision-making. For example, high performance ensures seamless interaction with large datasets, enabling rapid diagnosis and analysis.
- Accessibility: This parameter evaluates whether the modules are usable by individuals with different types of disabilities, including visual, auditory, motor, or cognitive impairments. The evaluation ensures that the modules comply with international accessibility standards, such as the web content accessibility guidelines (WCAGs), which are designed to make web applications more accessible to all users.
- Best Practices: This metric assesses the platform's compliance with modern web development standards, including security and resource optimization. In healthcare, where sensitive patient data are involved, adherence to these best practices ensures robust and reliable performance.
- Search Engine Optimization (SEO): SEO is used to ensure that the platform's content is well structured and easily indexable, contributing to the user experience, which aids in the dissemination of research findings and platform awareness among broader medical communities.

These evaluation criteria were chosen to comprehensively assess the visual representation capabilities of the DECODE platform, ensuring that it provides a reliable, accessible, and high-performing tool for clinicians and researchers involved in the diagnosis and treatment of PAD.

5. Results

The DECODE cloud platform was developed and evaluated to enhance the diagnosis, analysis, and treatment planning for PAD through its sophisticated visualization and data management capabilities. The platform integrates multiple modules, such as datasets, image processing, computational tools, segmentation tools, 3D reconstructions, and visual representations, each designed to handle specific aspects of medical imaging and data processing, as detailed in Figure 5.

The MPV module, depicted in Figure 6, was instrumental in facilitating the detailed exploration of 2D DICOM images across multiple planes. This module's advanced capabilities enable clinicians to obtain comprehensive views of the anatomical structures of the PAD. The implementation of a PWA provided seamless access across various devices, significantly enhancing the platform's usability.



Figure 6: Multiplanar Visualization of Medical Imaging Data in the DECODE Cloud Platform

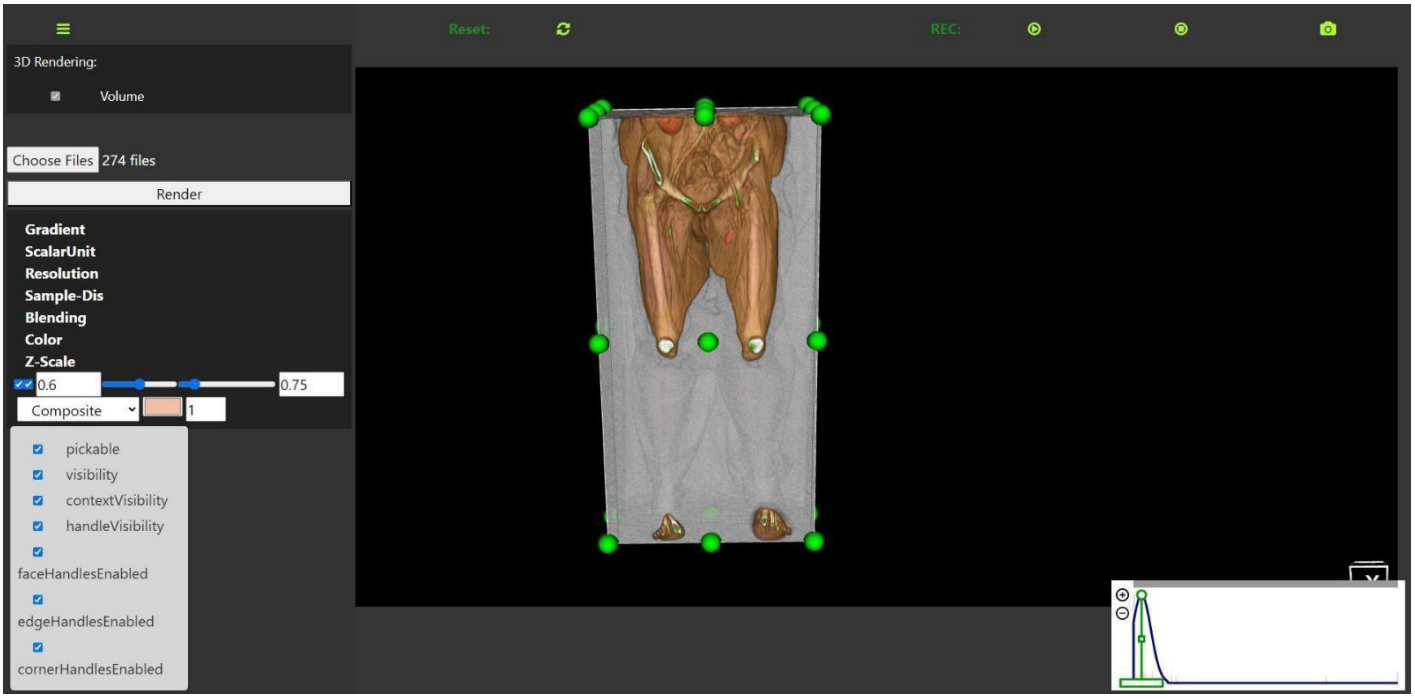


Figure 7: User Interface for 3D Rendering of Medical Imaging Data in the DECODE Cloud Platform.

The MPV module is accessible via the RESTful API endpoint/*api/visualization/mpv/*, which allows for efficient integration and data retrieval. This API endpoint supports various functionalities, including the collection of DICOM images, metadata retrieval, and real-time updates. The interactive tools for zooming, panning, and navigating through different planes (axial, sagittal, and coronal) are seamlessly integrated through this API, enabling the precise examination of specific regions of interest. These capabilities ensure that clinicians can interact with the visualization data dynamically, facilitating real-time adjustments and increasing diagnostic accuracy. The robust architecture of the MPV module, combined with its API-driven approach, ensures that the visualization tools are both scalable and adaptable to different clinical requirements. This integration not only improves the efficiency of data handling and processing but also ensures that the visualization remains consistent and reliable across various use cases and devices. Thus, the MPV module, with its advanced interface and RESTful API support, significantly enhances the diagnostic process and treatment planning for PAD.

The 3D rendering module, accessible via the RESTful API endpoint/*api/visualization/3d-rendering/*, is a critical component of the DECODE cloud platform. This endpoint facilitates efficient data retrieval and interaction, enabling real-time updates and dynamic adjustments. Figure 7 shows the user interface for this module, highlighting its robust functionality and interactive features that significantly enhance the analysis and treatment planning

Table 1: Performance Metrics for Multiplanar Visualization Module and 3D Rendering Module in the DECODE Cloud Platform obtained by Google's Lighthouse tool

Metrics (%)	Multiplanar Visualization Module	3D Rendering Module
Performance	94	89
Accessibility	90	90
Best Practices	100	100
Search Engine Optimization	100	90

for PAD. The user interface allows users to upload and render volumetric medical imaging data seamlessly. It includes a variety of controls for manipulating 3D visualization, such as gradient adjustment, scalar unit specification, resolution settings, and sample distance control. These controls enable users to tailor the rendering process to highlight specific anatomical features, facilitating a more precise examination of the imaging data.

Interactive features such as volume clipping, rotation, zooming, and panning are integrated into the interface, allowing users to explore the 3D models from various perspectives. The presence of checkboxes for enabling or disabling visibility, context visibility, handle visibility, and various handle controls (face, edge, corner) further enhances the flexibility and customization available to the user.

5.1. Evaluation of Visual Representation Modules

The performance metrics for the DECODE cloud platform's MPV module and 3D rendering module are summarized in Table 1. These metrics provide a quantitative assessment of the platform's capabilities across several key areas.

The MPV module demonstrated exceptional results, with a performance score 94%, indicating a highly responsive user experience. The module also scored 90% in accessibility, underscoring its commitment to inclusivity by meeting essential accessibility standards. In addition, the module achieved perfect scores (100%) in both best practices and SEO, reflecting its adherence to modern web development standards and its content's discoverability via search engines.

The 3D rendering module also performed strongly, with a performance score 89%. While slightly lower than the MPV module, this score still indicates a high level of efficiency, particularly given the complex computations required for 3D rendering. The module matched the MPV module with an accessibility score 90% and perfect scores 100% in both best practices and SEO, confirming its robust design and compliance with industry standards.

6. Discussion

The DECODE cloud platform represents a significant breakthrough in the diagnosis, analysis, and treatment planning for PAD. Leveraging cutting-edge cloud computing and advanced data visualization technologies, the DECODE platform surpasses the capabilities of current state-of-the-art (SoTA) solutions. It is the first platform specifically designed to integrate high-performance data processing, 3D medical imaging, and computational modeling into a unified system tailored for PAD. This integration allows for precise simulation of treatment interventions, such as drug-coated balloons, and provides clinicians with real-time, data-driven insights to optimize therapeutic strategies. Compared with other cloud-based platforms, such as teleophthalmology systems [39] and the BioData Catalyst (BDC) [40], DECODE has demonstrated significant advancements in terms of functionality and adaptability. While tele-ophthalmology systems excel in scalable image storage and annotation and BDC specializes in managing large-scale genomic datasets, DECODE uniquely integrates segmentation, 3D visualization, and real-time analysis to address PAD-specific clinical workflows. Furthermore, it mitigates common challenges observed in cloud-based solutions, such as inefficient handling of large datasets, access control, and vendor conformity, by employing tailored optimization techniques and ensuring robust security, scalability, and compliance with healthcare standards.

The DECODE platform overcomes the limitations of the SoTA by integrating comprehensive data management and advanced visualization tools. The MPV module's robust design and high performance underscore its effectiveness in medicine. The module's ability to provide detailed 2D and 3D views of anatomical structures significantly enhances diagnostic and treatment planning. The 3D rendering module's performance metrics also highlight its strengths in providing high-fidelity visualizations crucial for PAD treatment. The module's slightly lower performance score than the MPV module can be attributed to the complex computations required for rendering detailed 3D models. However, its

accessibility score and perfect scores in best practices and SEOs indicate that it remains an invaluable tool for clinicians.

While the performance metrics highlight the platform's robustness, several limitations must be acknowledged. One notable challenge is scalability, as the platform's current architecture may encounter bottlenecks when handling significantly larger datasets or accommodating simultaneous users over extended periods. Such scenarios could impact real-time processing and responsiveness, particularly in high-demand clinical settings. In addition, the reliance on high-performance GPUs and advanced cloud infrastructure poses a limitation, potentially restricting platform adoption in resource-constrained environments where access to such resources may be limited.

Future work will prioritize optimizing the platform architecture to increase scalability and explore cost-effective hardware alternatives, thereby ensuring broader accessibility and adoption in diverse healthcare settings. In addition, the planned development of a 3D segmentation and reconstruction module will focus on the precise delineation of anatomical structures, specifically targeting the segmentation of the peripheral artery with intima and adventitia thickness. This enhancement will enable detailed 3D reconstructions, improve the analysis of vascular health and support the personalization of treatment strategies. By accurately modeling these critical arterial layers, the module will provide clinicians with deeper insight into disease progression, plaque buildup, and potential treatment outcomes, ultimately contributing to more precise and effective clinical interventions. The integration of a virtual reality (VR) module will further enhance the platform by offering immersive visualization tools to improve spatial understanding of complex anatomical structures, supporting preoperative planning, educational applications, and patient engagement. Challenges such as ensuring hardware compatibility, optimizing real-time performance, and maintaining compliance with healthcare data standards will be actively addressed to maximize the impact of these innovations. These advancements position the DECODE platform as a versatile and comprehensive solution that not only surpasses existing SoTA solutions but also sets a new benchmark for cloud-based technologies in PAD diagnosis and management.

7. Conclusion

The DECODE cloud platform represents a transformative advancement in the diagnosis, analysis, and treatment of peripheral artery disease (PAD). By leveraging the capabilities of cloud computing and high-performance data visualization, the platform addresses the significant limitations of traditional diagnostic methods, offering a noninvasive, cost-effective, and scalable solution. The DECODE platform's comprehensive modules, including a data warehouse and multiplanar reconstruction with 2D and 3D rendering visualization, enable detailed and interactive processing and presentation of medical imaging data. The evaluation metrics underscore the platform's high performance and user accessibility. The multiplanar visualization module achieves a performance score 94, an accessibility score 90, and perfect scores 100 in best practices and search engine optimization (SEO). Similarly, the 3D rendering module scored 89 in performance, 90 in accessibility, and 100 in best practices and SEO, demonstrating the platform's robust design and compliance with industry standards. Our work highlights the transformative potential of integrating cloud computing in medical diagnostics, paving the way for further research and development in this critical area of healthcare.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956470, as part of the DECODE project.

References

[1] M. M. S. Radwan, Sini Thomas, Sithara, "Environmental Factors and Peripheral Artery Disease," in Environmental Factors in the Pathogenesis of Cardiovascular Diseases: Springer, 2024, 193-208.

[2] M. N. Søgaard, Peter Brønnum Eldrup, Nikolaj Behrendt, Christian-Alexander Nicolajsen, Chalotte W Lip, Gregory YH Skjøth, Flemming, European Journal of Vascular Endovascular Surgery, "Epidemiological trends and projections of incidence, prevalence, and disease related mortality associated with peripheral arterial disease: observations using nationwide Danish data," vol. 66, no. 5, 662-669, 2023, doi: 10.1016/j.ejvs.2023.08.005.

[3] S. P. S. Cartland, Christopher P Bursill, Christina Passam, Freda Figtree, Gemma A Patel, Sanjay Loa, Jacky Golledge, Jonathan Robinson, David A Aitken, Sarah, International Journal of Molecular Sciences, "Sex, Endothelial Cell Functions, and Peripheral Artery Disease," vol. 24, no. 24, 17439, 2023.

[4] S. Research, "Peripheral Artery Disease Market Growth, Trends and Forecast to 2031 " 2024, doi: 10.3390/ijms242417439.

[5] R. U. M. Center, "Peripheral Vascular Disease (PVD), https://www.rush.edu/conditions/peripheral-vascular- disease-pvd# " 2024

[6] J. D. Csore, Madeline Roy, Trisha L, Journal of Vascular Surgery Cases, Innovations Techniques, "Peripheral arterial disease treatment planning using noninvasive and invasive imaging methods," vol. 9, no. 4, 101263, 2023, doi: 10.1016/j.jvscit.2023.101263.

[7] W. R. Abbaoui, Sara El Bhiri, Brahim Kharmoum, Nassim Ziti, Soumia, Informatics in Medicine Unlocked, "Towards

revolutionizing precision healthcare: A systematic literature review of artificial intelligence methods in precision medicine," 101475, 2024, doi: 10.1016/j.imu.2024.101475.

[8] G. G. Gabrani, Sunil Vyas, Sonali Arya, Pradeep, "Revolutionizing Healthcare: Impact of Artificial Intelligence in Disease Diagnosis, Treatment, and Patient Care," in Handbook on Augmenting Telehealth Services: CRC Press, 2024, 17-31.

[9] K. K. Kavitha, C, "Cloud-Based Data Analytics for Healthcare 5.0," in Pioneering Smart Healthcare 5.0 with IoT, Federated Learning, and Cloud Security: IGI Global, 2024, 44-56, doi: 10.4018/979-8-3693-2639-8.ch004.

[10] A. A. Shalan, Abubakar Habli, Ibrahim Tew, Garry Thompson, Andrew, "YORwalK: desiging a smartphone exercise application for people with intermittent claudication," in Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth: IOS Press, 2018, 311-315, doi: 10.3233/978-1-61499-852-5-311.

[11] K. S. Paldán, Jan Ullrich, Greta Steinmetz, Martin Rammos, Christos Jánosi, Rolf Alexander Moebus, Susanne Rassaf, Tienush Lortz, Julia, JMIR research protocols, "Feasibility and clinical relevance of a mobile intervention using TrackPAD to support supervised exercise therapy in patients with peripheral arterial disease: study protocol for a randomized controlled pilot trial," vol. 8, no. 6, e13651, 2019.

[12] A. V. Harzand, Alexander A Alrohaibani, Alaaeddin Abdelhamid, Smah M Gordon, Neil F Thiel, John Benarroch - Gampel, Jaime Teodorescu, Victoria J Minton, Keri Wenger, Nanette K, Clinical Cardiology, "Rationale and design of a smartphone - enabled, home - based exercise program in patients with symptomatic peripheral arterial disease: the smart step randomized trial," vol. 43, no. 6, 37-545, 2020, doi: 10.1002/clc.23362.

[13] J. S. Lortz, Jan Kuether, Tabea Kreitschmann-Andermahr, Ilonka Ullrich, Greta Steinmetz, Martin Rammos, Christos Jánosi, Rolf Alexander Moebus, Susanne Rassaf, Tienush, JMIR Formative Research, "Needs and requirements in the designing of mobile interventions for patients with peripheral arterial disease: questionnaire study," vol. 4, no. 8, e15669, 2020.

[14] H. P. Paredes, Dennis Barroso, João Abrantes, Catarina Machado, Isabel Silva, Ivone, "Supervised physical exercise therapy of peripheral artery disease patients: M-health challenges and opportunities," 2021, http://hdl.handle.net/10125/71086.

[15] A. M. D. Flores, Falen Leeper, Nicholas J Ross, Elsie Gyang, Circulation research, "Leveraging machine learning and artificial intelligence to improve peripheral artery disease detection, treatment, and outcomes," vol. 128, no. 12, 1833-1850, 2021, doi: 10.1161/CIRCRESAHA.121.31822.

[16] N. M. Forghani, Keivan Dabanloo, Nader Jafarnia Farahani, Ali Vasheghani Forouzanfar, Mohamad, IEEE Journal of Biomedical Health Informatics, "Intelligent oscillometric system for automatic detection of peripheral arterial disease," vol. 25, no. 8, 3209-3218, 2021, doi: 10.1109/JBHI.2021.3065379.

[17] T. G. Collins, Mugur Overton, Kathryn Benton, Mary Lu, Liuqiang Khan, Faarina Rohleder, Mason Ahluwalia, Jasjit Resnicow, Ken Zhu, Yiliang, JMIR Formative Research, "Use of a smartphone app versus motivational interviewing to increase walking distance and weight loss in overweight/obese adults with peripheral artery disease: pilot randomized trial," vol. 6, no. 2, e30295, 2022.

[18] M. K. Kim, Yesol Choi, Mona, BMC Medical Informatics Decision Making, "Mobile health platform based on user-centered design to promote exercise for patients with peripheral artery disease," vol. 22, no. 1, 206, 2022, doi: 10.1186/s12911-022-01945-z.

[19] F.-Q. D. Wu, Qian-Wan Wang, Ji-Guang Li, Wen-Zhu, Current Treatment Options in Cardiovascular Medicine, "The Role of Supervised Exercise Therapy in the Management of Symptomatic Peripheral Artery Disease with Intermittent Claudication," vol. 25, no. 10, 501-513, 2023, doi: 10.1007/s11936-023-01001-7.

[20] J. I. Zaki, SM Riazul Alghamdi, Norah Saleh Abdullah-Al-Wadud, Mohammad Kwak, Kyung-Sup, IEEE Access, "Introducing cloud-assisted micro-service-based software development framework for healthcare systems," vol. 10, 33332-33348, 2022, doi: 10.1109/ACCESS.2022.3161455.

[21] D. O. Alekseeva, Aleksandr Arponen, Otso Lohan, Elena Simona, Computer Science Review, "The future of computing paradigms for medical and emergency applications," vol. 45, p. 100494, 2022, doi: 10.1016/j.cosrev.2022.100494.

[22] M.-A. B. Filz, Jan Philipp Herrmann, Christoph, Journal of Intelligent Manufacturing, "Digitalization platform for data-driven quality management in multi-stage manufacturing systems," vol. 35, no. 6, 2699-2718, 2024, doi: 10.1007/s10845-023-02162-9.

[23] AboArab, Mohammed A., Vassiliki T. Potsika, Nikola Petrović, and Dimitrios I. Fotiadis, "DECODE cloud platform: A new cloud platform to combat the burden of peripheral artery disease," in 2022 Panhellenic Conference on Electronics & Telecommunications (PACET), 2022, 1-6: IEEE, doi: 10.1109/PACET56979.2022.9976356.

[24] Z. A. K. Mughal, Adnan Ahmed, Syed Sohail Qazi, Salman, Journal of Software Engineering, "Key factors and features analysis of popular SaaS ERP Systems for Adoptability," vol. 1, no. 1, 11-21, 2022.

[25] D. C. L. De Oliveira, Ji Pacitti, Esther, Data-Intensive Workflow Management. Springer Nature, 2022.

[26] R. K. Fauzan, Ice Nurwibowo, Bima Dinda Wibowo, Della Aulia, IPTEK The Journal for Technology Science, "A systematic literature review on progressive web application practice and challenges," vol. 33, no. 1, 43-58, 2022, doi: 10.12962/j20882033.v33i1.13904.

[27] D. V. Hesmondhalgh, Raquel Campos Kaye, D Bondy Valdovinos Li, Zhongwei, Media Communication, "Digital platforms and infrastructure in the realm of culture," vol. 11, no. 2, 296-306, 2023, doi: 10.17645/mac.v11i2.6422.

[28] A. G. Blum, R Rauch, A Urbaneja, A Biouichi, H Dodin, G Germain, E Lombard, C Jaquet, P Louis, M, Diagnostic interventional imaging, "3D reconstructions, 4D imaging and postprocessing with CT in musculoskeletal disorders: past, present and future," vol. 101, no. 11, 693-705, 2020, doi: 10.1016/j.diii.2020.09.008.

[29] M. T. Thakkar, Mohit, Building React Apps with Server-Side Rendering: Use React, Redux, Next to Build Full Server-Side Rendering Applications, "Introducing react. js," 41-91, 2020.

[30] E. U. Ziegler, Trinity Brown, Danny Petts, James Pieper, Steve D Lewis, Rob Hafey, Chris Harris, Gordon J, JCO clinical cancer informatics, "Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research," vol. 4, 336-345, 2020, doi: 10.1200/CCI.19.00131.

[31] F. V. Paligu, Cihan, Future Internet, "Browser Forensic Investigations of Instagram Utilizing IndexedDB Persistent Storage," vol. 14, no. 6, p. 188, 2022, doi: 10.3390/fi14060188.

[32] J. K. Udupa, "3D imaging: principles and approaches," in 3D Imaging in Medicine, Second Edition: CRC Press, 2023, 1-73.

[33] R. L. T. Cieri, Morgan L Carney, Ryan M Falkingham, Peter L Kirk, Alexander M Wang, Tobias Jensen, Bjarke Novotny, Johannes Tveite, Joshua Gatesy, Stephen M, Journal of morphology, "Virtual and augmented reality: New tools for visualizing, analyzing, and communicating complex morphology," vol. 282, no. 12, 1785-1800, 2021, doi: 10.1002/jmor.21421.

[34] P. J. Thakur, Prashant, "Django: Developing web using Python," in 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2023, pp. 303-306: IEEE, doi: 10.1109/ICACITE57410.2023.10183246.

[35] Kitware, Inc. (2024). VTK.js. Accessed July 2024. https://kitware.github.io/vtk-js, doi: 10.5281/zenodo.4552412, 2021.

[36] S. J. Matt McCormick, Forrest Li, Alexis Girault, David Thompson, Juan Carlos Prieto, Scott Wittenburg, & HastingsGreer. (2021). InsightSoftwareConsortium/itk-js: v14.1.1 (v14.1.1), doi: 10.5281/zenodo.4957207.

[37] R. JohnsonChris, Health Data Science, "A review of three-dimensional medical image visualization," 2022, doi: 10.34133/2022/9840519.

[38] M. P. Manca, Vanessa Paternò, Fabio Santoro, Carmen, ACM Transactions on Accessible Computing, "The transparency of automatic web accessibility evaluation tools: design criteria, state of the art, and user perception," vol. 16, no. 1, 1-36, 2023, doi: 10.1145/3556979.

[39] Schweitzer, M., Ostheimer, P., Lins, A., Romano, V., Steger, B., Baumgarten, D., & Augustin, M. (2024). Transforming Tele-Ophthalmology: Utilizing Cloud Computing for Remote Eye Care. In dHealth 2024 (pp. 215-220). IOS Press, doi: 10.3233/SHTI240040.

[40] Ahalt, S., Avillach, P., Boyles, R., Bradford, K., Cox, S., Davis-Dusenbery, B., ... & Asare, J. (2023). Building a collaborative cloud platform to accelerate heart, lung, blood, and sleep research. Journal of the American Medical Informatics Association, 30(7), 1293-1300, doi: 10.1093/jamia/ocad048.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).

Received: 15 September, 2024 July, 2024, Revised: 30 October, 2024, Accepted: 23 November, 2024, Online: 22 December, 2024

DOI: https://doi.org/10.55708/js0312002

Advanced Digital Twin of a Industrial Robotic System for Measuring Pipe Wall Thickness in Nuclear Power Plants

Rogério Adas Pereira Vitalli *, João Manoel Losada Moreira

Federal University of ABC (UFABC), Interdisciplinary Laboratory of Nuclear Energy (NUC-LAB), Santo André, São Paulo, 09210-580, Brazil

*Corresponding author: Rogério Adas Pereira Vitalli, rogerio.vitalli@ufabc.edu.br

ABSTRACT: This paper presents the development of the digital twin of an advanced industrial robotic system for pipe wall thickness inspection in the turbine building of nuclear power plants. The robotic inspection system consists of 3 units, the first being the mobile unit, the second a robotic unit for automatic pipe wall thickness measurement using the ultrasound technique, and the third unit for power supply, communication and auxiliary services. To develop the advanced robotic inspection system, a prototype was built in the laboratory to study different situations and geometric configurations that may arise in the field (turbine building). With Process Simulate software version 16.1.2, the digital twin of the prototype was developed including industrial robot, metal platform and pipe sections. This paper presents the results for the virtual commissioning of the wall thickness measurement of a vertical pipe section and the novel design of the End-Effector of the industrial robot. Therefore, technical discussions are made on the requirements to deal from the design to the functional requirements of developing the end-effector, digital twin and the creation of a commissioning method for industrial robots. The analyses and insights from virtual commissioning made it possible to verify that the robot could access any part of the pipe surface through interpolated movements with spatial coordinates of the robotic arm along its sides.

KEYWORDS: Digital Twin, Ultrasound Measurement, Pipe Wall Thickness, Virtual Commissioning, Process Simulate

1. Introduction

Complex industrial facilities such as nuclear power plants require pipe wall thickness inspection for safe and reliable operation. These pipes, submitted to high pressure and temperature, connect the steam generators, turbine, condenser, and may undergo thickness reduction due to corrosion and erosion [1]-[3]. The inspection environment is hot and humid (50ºC and 100 % relative humidity). The field space has complicated geometry and access to inspection locations, the pipes have diameters varying from 50 cm to 65 cm, and are positioned in horizontal, vertical, or inclined directions. During plant panned shutdown, the inspections may include tens of thousands measurement points of wall thickness [4], [5].

Traditional inspection systems are usually rigid and require intense human action to circumvent field unforeseen scenarios to make wall thickness measurement. Given the amount of work involved during inspections, it is desirable to automate these activities, and foresee access difficulties and variability of inspection conditions and pipes geometric configurations. To increase the efficiency to handle complex activities in changing scenarios, several authors propose the use of robotic automation systems with and without user assistive actions [6]-[8], and the use of digital twins and other industry 4.0 approaches to streamline their design and commissioning phases [9].

This work presents part of the development of a robotic system for measuring the wall thickness of the secondary system pipes of the Angra 1 nuclear power plant, more specifically the development of its digital twin. This method allows dividing the development process into two stages, the first being development via simulation (virtual project) and the second stage in which the physical project is built from the virtual project data, including the robot

programming that is exported from the virtual project to the field robot [10].

Digital twins are very realistic virtual models of a process and its interaction with the actual world environment [11]. They include the equipment and all the steps to carry out a certain production process [12]. Using the benefits of the digital twin and virtual commissioning tools, it is possible to increase communication and coordination between the parts of the process, allowing for smarter decision-making [12]. This technology makes it possible to quickly and flexibly develop different situations aimed at automating and integrating operator agents and robes to perform complex tasks [13]. It allows integrating sensory data from physical assets and simulating a variety of robotic conditions in which an agent can interact. As a result, the agent can learn an adaptive task allocation strategy that increases project performance. In [13], the authors tested this method for the robot-automated construction problem. Results indicated that model task allocation reduced build time by 36% in three dynamic test environments when compared to the conventional method.

The biggest benefit of virtual commissioning is the savings. It is possible to test an investment in advance, which could be hundreds of thousands or even millions of dollars in equipment, since a single robot could cost you around $50 to $70,000. [14]. So, five robots, equipment, clamp systems, accessories, different products, and all that together in one manufacturing cell is a significant investment. Companies do not intend to develop projects without knowing if it can satisfactorily perform at the expected levels and results. Process reduction Standard operating. Possibility of carry out part of this process in a more convenient environment (no necessarily on-site) combined with the opportunity to use the emulation model for training workers. Parallel development and optimization of mechanical parts, especially mechanisms mechatronics. [15]. Programming and debugging simultaneously from the control software. The Companies aim to implement sustainable manufacturing of robotic cells, in order to improve profitability, reducing resource consumption, global expenses, as well as satisfying the regulatory input of ecological impacts. In addition, the widespread adoption of industrial robots, necessary to satisfy the ever-increasing requirements in terms of manufacturing quality, customization and flexibility, further increased the need to improve the energy efficiency of robotic cells. [16].

This article is organized as follows: section 2 presents a State of Art, section 3 presents a brief description of the robotic system for pipe wall thickness measurement and section 4 presents the method used in this research, data and details of the equipment used. In section 5 we present the results and discussions and in section 6, the final considerations.

2. State of Art

This paper seeks to present, through emerging projects, the concept of digital twins was named one of Gartner's Top 10 Strategic Technologies Trends for 2022 [17]. Thomas Kaiser, SAP Senior Vice President, IoT, put it this way: “Digital twins are becoming a business imperative, covering the entire lifecycle of an asset or process and forming the basis for connected products and services. Companies that fail to respond will be left behind.” It is estimated that within three to five years, billions of things will be represented by digital twins [18]. Using physical data about how a thing's components operate and respond to the environment, as well as data provided by sensors in the physical world, a digital twin can be used to analyze and simulate real-world conditions, respond to changes, improve operations, and add value. Industry 4.0 (i4.0), which has been known as the “Fourth Industrial Revolution”, emerges with the aim of meeting this new global demand with customized products and services [19]. One of the concepts that will be widely used in i4.0 is the Digital Twin, which from a production perspective, incorporates the virtual context into the real context of a production system. In [20], Digital Twins are very realistic virtual models of the current state of the process and its own behavior in interaction with the real-world environment, including the equipment and all the steps to carry out a given production process [21], [22]. Using the benefits of the digital twin (DT) and virtual commissioning (VC) tools, it is possible to increase the advantage of communication and coordination between manufacturing areas, allowing for smarter decision-making. This allows automated systems and robotic cells to start operating (running) much faster and with fewer errors [23].

According to [24], the growing use of Virtual Commissioning during the development process of automated factories, combined with the growing demand for better quality control, leads to the need for improved virtual plants that systematize the configuration procedures required to perform their processes. Common plant simulation techniques based on the concept of virtual commissioning of robotic cells go beyond the need to validate control algorithms. In other words, new approaches need to be developed to meet the demand for reconfiguring their operational resources in a systematic way, compatible with the flexibility that these autonomous resources currently have. The need for new solutions in Robotics Engineering® for the design of complex projects involving physical systems and the virtual part associated with them has never been so present. Virtual commissioning technology can be considered as one of the established trends in automotive

assembly. Among other benefits, it promotes more efficient treatment of the complexity associated with assembly systems, capable of causing a reduction in the acceleration time of the system itself and a reduction in the development time of the product capable of meeting market competitiveness.

According to [25], the digital twin synthesizes sensory data from physical assets and is used to simulate a variety of robotic construction site conditions in which an agent may interact. As a result, the agent can learn an adaptive task allocation strategy that increases project performance. They tested this method with a case project in which a virtual robotic construction project (i.e., interlocking concrete bricks are delivered and assembled by robots) was digitally twinned for training. The results indicated that the model task allocation reduced construction time by 36% in three dynamic test environments when compared to an imperative model rule-based method.

Along with the trends of mass customization, flexible robotic applications are becoming increasingly popular. Although conventional robotic automation of workpiece handling seems to have been solved, advanced tasks still require a lot of effort to achieve. In most cases, on-site robot programming methods, which are intuitive and easy to use, are not applicable in flexible scenarios. On the other hand, the application of offline programming methods requires careful modeling and planning. Consequently, [26] proposes a generalized approach to the development methodology for flexible pick-and-place robotic work cells in order to provide guidance and thus facilitate the development process. The methodology is based on the Digital Twin (DT) concept, which allows iterative refinement of the work cell in both the digital and physical space. The goal is to streamline the overall commissioning (or reconfiguration) process and reduce the amount of work in the physical work cell. This can be achieved by digitizing and automating the development and maintaining sufficient proximity. With this, the operation of the digital model can be performed accurately in the physical work cell. The methodology is presented through a semi-structured pick-and-place task, performed in an experimental robotic work cell in conjunction with a reconfiguration scenario [27]-[28].

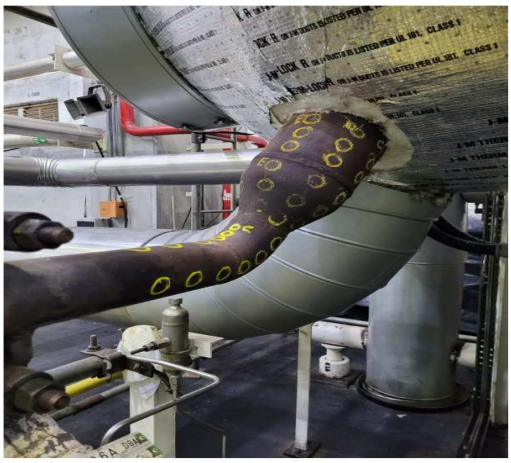
3. Brief Description on of the Advanced Robotics System

The Robotic System for pipe wall Thickness Measurement (RSTM) aims to meet the licensing needs of the Angra 1 Nuclear Power Plant for long-term operation in fast and efficient manner [29]. This RSTM has 3 units, the first, named Unit of Remotely Operated Vehicle (UROV), being a vehicle that can be operated remotely to allow displacement of the whole RSTM to the pipe area under inspection. The second unit, named Unit of Measuring Robotic Cell (UMRC), is a robotic work cell indexed on the UROV and equipped with a robotic arm that performs the movements and inspection actions and an effector with sensors for spatial orientation and for measuring pipe wall thickness through the ultrasound technique. The third unit contains all other RSTM systems such as power supply cables, data transmission cables, real-time information, computer, and ancillary equipment. This unit is named Unit of Communication, Power supply and Ancillary services (UCPA) [29].

The RSTM must be able to approach the defined inspection location and measure the thickness of the pipe wall. Figure 1 shows the complexity and geometric variability of the piping system in the field (turbine building). The solution provided for the inspection problem is a hybrid, contemplating the assistance action of the user to approach the inspection region, and robotic automation to perform the final approach to the inspection locations on the pipe surface and wall thickness measurements. An example of actual inspection locations is the grid of yellow circles shown in Fig. 1. The robot end-effector has an ultrasound-based approximation sensor to provide spatial orientation and an ultrasound-based sensor for pipe wall thickness measurement [30].

Figure 1: Piping section showing the environment and situation of pipe inspection at Angra nuclear power plant 1. Circles show positions indicated by the engineering group for pipe thickness inspection.

The user operates the UROV to approach the RSTM to



the measuring locations on the pipe external surface. The first grid location must be manually identified and recorded by the user, that is, with the help of the "teach pendant" he or she moves the industrial robot end-effector through the piping environment to the first measurement position marked on the pipe wall and register its coordinates. From there, the robot present in the UMRC, with its spatial orientation system, automatically scans all the measuring points on the grid and automatically performs the wall thickness measurements. After carrying out the measurement, the result must be transmitted to the UCPA through the communication interface with the plant.

While conventional robotic automation for handling workpieces and performing specific tasks seems to have been resolved, more complex tasks in less controlled environments, such as in the field, require more flexible solutions. In most cases, robot programming methods in a controlled environment are not applicable in variable scenarios found in the real world [31].

4. Methods, Data and Equipment

The use of the Digital Twin concept in an iterative way, both in the digital space and in the physical space, to refine the work cell development process is the solution adopted in this work. The objective is to streamline the overall commissioning process with different configurations or reconfigurations according to the work situation [31]-[32].

Figure 2 presents the main steps of the approach based on the digital twin technique for the UMRC development. The software adopted to develop the digital twin is the Tecnomatix Process Simulate, version 16.1.2 and the robot used is the ABB IRB 1600-6/1.45 model.

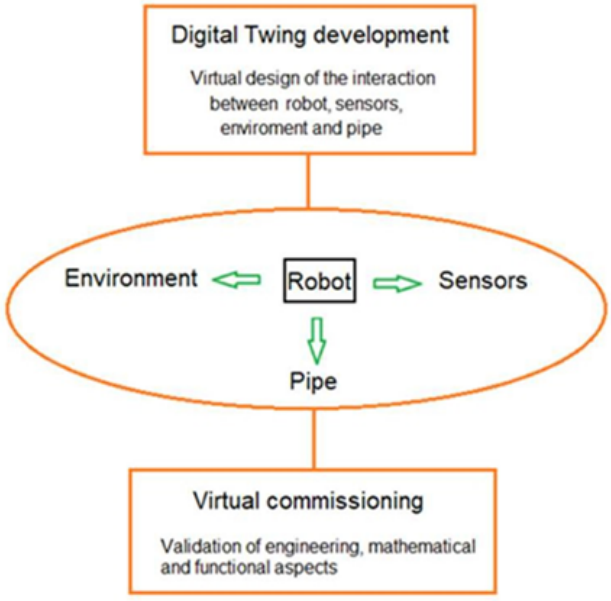


Figure 2: Scheme of the method for developing robotic systems using the digital twin technique. All the capabilities of the robot are developed via programming (software) and virtual commissioning is carried out via simulation of the various tasks that the robot must perform until its validation.

Figure 3 shows the robotic cell prototype installed at the UFABC Robotics Laboratory to develop the UMRC that contains the intelligent part of the robotic measurement system. The robotic cell includes ABB IRB 1600-6/1.45 robot, the metal platform and two pipe sections taken from the Angra 1 nuclear power plant. The pipe section is assembled on the platform in different geometric configurations and situations emulating actual measurement conditions. These configurations are obtained from field situations like the one shown in Figure 3 or others from field situations as shown in Figure 1.

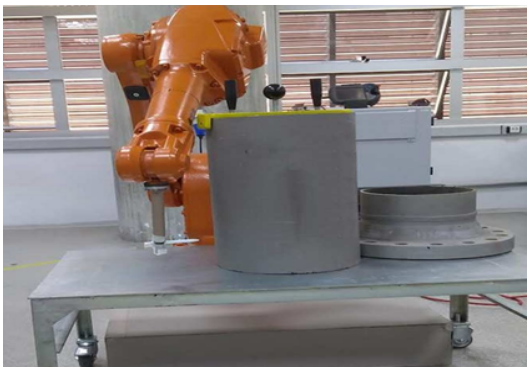


Figure 3: Piping Robotic cell in the Laboratory emulating the UMRC. One sees the ABB IRB 1600-6/1.45 robot, the metallic platform and two pipe sections taken from the Angra 1 power plant

Automatic measurement was performed numerous times using various measurement techniques to ensure repeatability, accuracy and robustness of the measured data. The ultrasonic thickness gauge from the manufacturer METROTOKYO used was the MTK-1301 PRIME model. The electronic transducer was the standard 5MHz/Ø10mm model. The longitudinal wave speed for the "glycerin" material was 1920 m/s and for the 330-steel material it was 5600 m/s, respectively, the speed of sound. Figure 4 shows the specification of the system's measuring sensors.



Figure 4: Meters used: A-) Model Prime 1300. B-) Model Prime 1301. C-) Transducer Model.

The older model did not allow data extraction, as it did not have a USB connection. Therefore, the MTK-1301 PRIME meter model was used, with the following features:

- Manual or automatic shutdown after a period of non-use;
- Table of 9 pre-established ultrasound speeds;
- Calibration function with 3mm standard attached to the meter;
- Calibration of the speed of sound with 1 or 2 points;
- Simple or sweep measurement;
- Individual measurement or with average calculation;
- Measurement with audible alarm for minimum/maximum tolerance;
- Measurement with differential calculation with percentage;
- Measurement on a surface with high temperature (optional sensor);

- Improved measurement for pipe walls. measurement

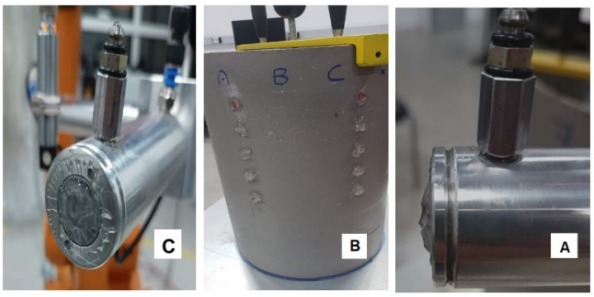


Figure 5: Ideal amount of glycerin: A-) Approximately 50g of glycerin controlled by the time the regulating valve is activated. B-) Spraying at 40 different points in the piping. C-) Visual inspection of the "micro-mesh" to prevent drying.

Figure 5 shows evidence of the unprecedented procedure developed in the laboratory for applying the coupling liquid. It was observed through numerous tests that a small amount of coupling liquid (glycerin) interferes with the quality of the measurement, that is, there are chemical reactions between the electronic transducer nozzle, the type of pipe material and the amount of glycerin. The measurements remained stable with two sprays per point and the robotic manipulator program was adjusted for this function through the "pulse" command.

The Communication, Power Supply and Ancillary Services Unit (UCPA) fulfilled the scope of its title. For the communication part, the principles of Industry 4.0 were adopted, which seeks to integrate the "work field", the engineering group and other sectors that Eletronuclear may require. After the thickness measurement was performed, the result was transmitted to the UCPA through a communication interface. These enabling technologies of Industry 4.0 follow data security and interface communication protocols based on APIs (Application Programming Interface) that allow applications to be customized according to Eletronuclear's needs. The UCPA was created and developed based on a new, more modern thickness gauge model that allowed access and transfer of the measured data that was stored inside the gauge to an external computer. Figure 7 and Figure 7 shows both systems.

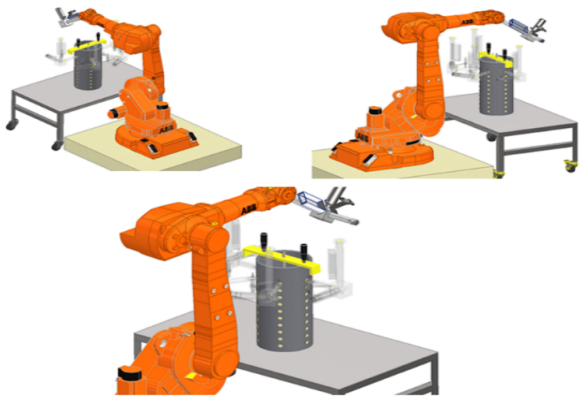


Figure 6: Different views and settings of automatic measurement.



Figure 7: UCPA Communication Interface.

5. Results and Discussions

The workstation geometric data representing the UMRC prototype (Figure 3), containing information about the robot, workstation platform, piping section and sensors, were transferred to the Process Simulate software. After modeling the UMRC prototype digital twin with all its physical characteristics (ABB IRB 1600-6/1.45 robot and end- effector), mathematics, kinematics, and information from the environment, virtual commissioning began in order to test and validate the prototype solution for building the robotic cell. Figure 4 presents the UMRC prototype digital twin developed with Process Simulate software for the configuration of a pipe section with vertical orientation. The two boxes connected at the end of the robotic arm represent volumes occupied by the end-effector. Table 1 presents the positioning of the robotic arm for a thickness measurement activity performed during virtual commissioning. It shows the positioning in terms of the angles from the 6 joints of the robot's digital twin at the measurement location indicated in Figure 8.

Virtual commissioning was performed for various locations of wall thickness measurement in this vertical pipe section ensuring that the robot could accesses any part of the pipe external surface and perform measurements. Accesses were verified by moving the robotic arm along the sides of the pipe. The result in Table I exemplifies a measurement location accessed by the robot's digital twin.

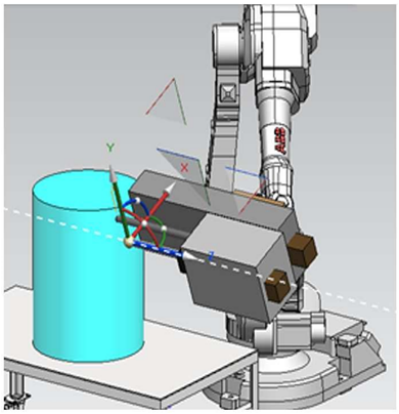


Figure 8: Digital twin of ABB model IRB 1600-6/1.45 robotic manipulator on workstation representing UMRC unit. The digital twin shows the robot, pipe section, and end-effector frame. Digital validation was performed for various measurement configurations of pipe wall thickness.

Table 1: Angles for the joints furnished by the digital twin of robot ABB IRB 1600-6/1.45. The joint angles represent a measuring position on the pipe section shown in Figure 8.

Joint	Angle (degree)
j1	37.45
j2	2.98
j3	37.70
j4	71.51
j5	36.32
j6	141.12

Figure 8a presents the OLP program generated by Process Simulate for export to the real robot ABB IRB 1600-6/1.45.

```
VERSION:1
LANGUAGE: ENGLISH
MODULE irb120_3_0_58_PNP_Angra1
!# Process Simulate Disconnected 15.1.2 to ABB Rapid download
!# ----------
!# CREATED : 22/04/2023 16:31:22 by VITALLI
!# CELL : UEABC_ABB
!# ROB  irb1600_6_145
!# ------
!# ------ ROBTARGET
!# ----------
 LOCAL CONST robtarget Ent:=[[623.32,231.38,364.23],[0.659024,-
0.256295,0.659024,0.256295],[0,-1,-
4,1],[9E+09,9E+09,9E+09,9E+09,9E+09,9E+09]];
 LOCAL CONST robtarget pick1:=[[415.39,131.33,19.98],[0.659024,-
0.256295,0.659024,0.256295],[-1,-1,-1,-
1],[9E+09,9E+09,9E+09,9E+09,9E+09,9E+09]];
 LOCAL CONST robtarget Out:=[[415.39,131.33,104.73],[0.659024,-
0.256295,0.659024,0.256295],[0,-1,-
4,1],[9E+09,9E+09,9E+09,9E+09,9E+09,9E+09]];
 LOCAL CONST robtarget place1:=[[311.32,302.111,46.01],[0.659024,-
0.256295,0.659024,0.256295],[-1,-1,-1,-
1],[9E+09,9E+09,9E+09,9E+09,9E+09,9E+09]];
 LOCAL CONST robtarget
In_Place:=[[325.32,337.69,137.73],[0.659024,-
0.256295,0.659024,0.256295],[0,0,-
1,1],[9E+09,9E+09,9E+09,9E+09,9E+09,9E+09]];
 LOCAL CONST robtarget
Out_Place:=[[378.46,324.44,192.45],[0.659024,-
0.256295,0.659024,0.256295],[0,0,-
1,1],[9E+09,9E+09,9E+09,9E+09,9E+09,9E+09]];
ENDMODULE
```

Figure 8: *Offline* (OLP) robotic program generated by Process Simulate for export to the real ABB IRB 1600-6/1.45 robot. The program was generated in the ABB RAPID language for ABB industrial robots.

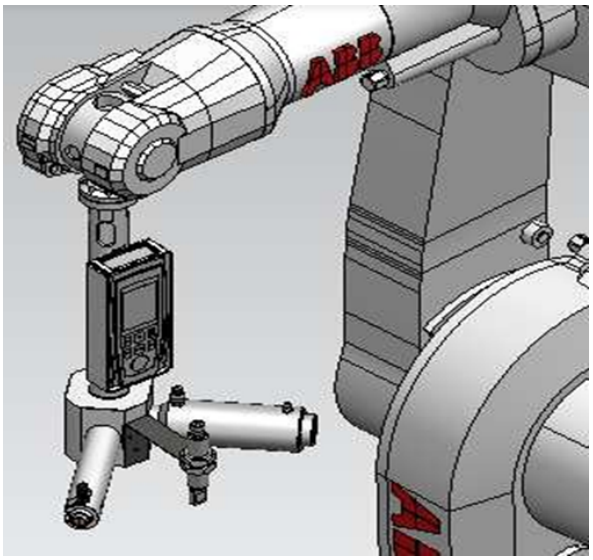


Figure 9: Final innovation design of the End-Effector.

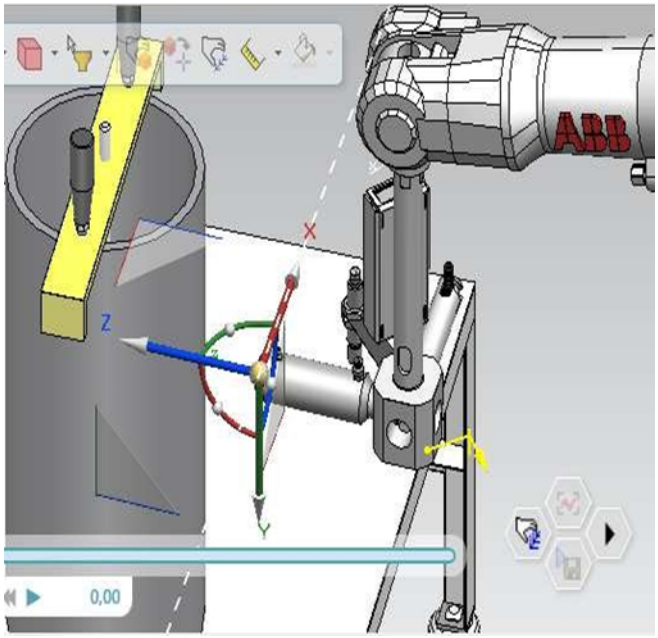


Figure 10: Projection of point vectors and spatial coordinates. Defined type of movement for the robot.

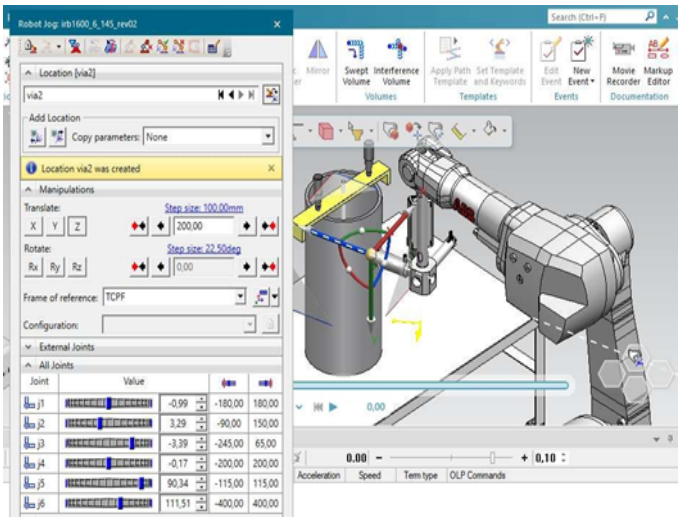


Figure 11: Recording of the first points of the trajectory of the ABB robotic manipulator with the Digital Twin scenario completed during automatic thickness measurement.



Figure 12: The ABC mesh on one side (right) of the pipe and the DEF mesh on the other side (left) of the pipe, together total 180 degrees of automatic measurement with the robot, divided into 10 layers with 40 distinct points.

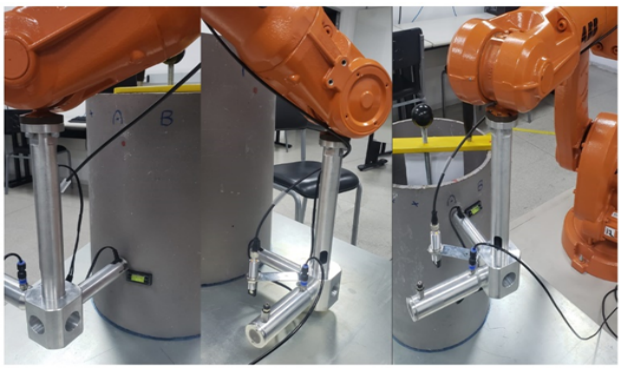


Figure 13: Automatic measurements on the last layer performed by the robotic system.

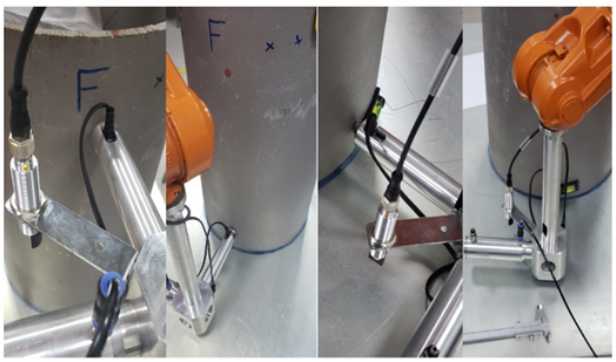


Figure 14: Automatic measurements on the first layer performed by the robotic system.



Figure 15: Example of automatic measurement performed by the robot in an unprecedented way with the value shown on the display.

Upload time:2024-6-21 23:15:52

Group 0001	No.	Thickness(mm)
	01	11,08
	02	10,43
	03	10,46
	04	10,55
	05	11,22
	06	10,55
	07	10,38
	08	10,62
	09	10,66
	10	10,79

Figure 16: Thickness measurement data from the first ten points measured on the pipe are automatically exported to any device via USB cable in a secure manner to avoid manual typing errors.

The success of automatic measurements with robotic technology was guaranteed by the innovation in the development of the automatic "spraying" process of the coupling liquid (glycerin). This accessory present in the robot's End-Effector was essential to spray the ideal amount of glycerin before the electronic transducer accurately measured the thickness at the point. It was configured and activated by the robot's control logic through a solenoid valve in conjunction with a unidirectional flow regulator to automatically control the output pressure within a range of 2.4 bar (Kgf/cm2). The ultrasonic sensor that accurately measures the distance of proximity of objects performed the measurement through a high-frequency sound wave, which is reflected by the object to be detected, thus avoiding collision between the robot and the pipe wall. The performance of the coupling liquid reservoir was extensively tested after corrections and improvements, guaranteeing expressive and unprecedented results. The End-Effector technology was 100% innovative, from its conception, design, improvements, manufacturing and implementation, validated in a real laboratory environment. The results of the digital twins revealed the importance and support they offer in the development of scenarios for the study of robotic applications. It should be noted that they were not used for real-time monitoring of the problem. The aerial trajectory positions for the robot and their respective corrections and adjustments were defined. The complete version of the digital twin was validated by comparing the virtual scenario with the real scenario. The robot's TCP was accurate and stable during the recording of points in a virtual environment. The first point was successfully recorded following the sequence of defined operations and without any type of human intervention.

The first complete thickness measurement operation was created in a customized virtual environment and as a result the OLP (offline program) code for the real industrial robot was generated. The entire laboratory layout was calibrated with a laser system to avoid variation of the piping on the bench in relation to the floor and the base of the robot. Numerous tests were performed on the robotic system before the measurements were taken, which were extensively tested.

The counter-tests of the measurements at the first points of the pipeline were performed numerous times using a caliper to ensure a standard for comparison. The discussion of physical contact in the pipeline for automatic thickness measurement between the electronic measuring transducer and the pipeline took place after spraying the coupling liquid; a fundamental factor for the measurement to occur. When the robot's End-Effector physically "touches" the pipeline, the sensor detects its presence and it enters "Hold" (a temporary stop). After this, the manipulator still exerts a small axial force to completely touch the face of the measuring transducer and ensure the quality of the procedure. Finally, the digital LCD display shows the measurement data and automatically records the value in its internal memory. Regarding the MTK 1301 Prime meter, the greatest difficulty was to equate the following variables: pre-established ultrasound speeds, calibration function with 3mm standard attached to the meter, sound speed calibration with 1 or 2 points, simple measurement or with sweep, individual measurement or with average calculation, measurement with audible alarm of minimum/maximum tolerance, measurement with differential calculation with percentage and improved measurement for pipe walls. The robotic manipulator operated at a safety speed of 10% (low inertia) in relation to the final nominal speed because the base of the robot was not "anchored" to the floor. The complete cycle time for spraying the glycerin and measuring each individual point took approximately 30 seconds.

6. Final Considerations

In this paper, we present the development of an advanced industrial robotic system for automatic inspection of tube wall thickness in the turbine building of nuclear power plants. A prototype of the system was built in the laboratory to study different situations and geometric configurations for measuring the external tube wall thickness. The digital twin of this prototype was developed using Process Simulate software version 16.1.2., including a 6-degree-of-freedom industrial robot, a metal platform, and nuclear tube sections. Digital validation was performed for various configurations of measuring the external tube wall thickness through virtual commissioning. The results of the virtual commissioning for a vertical pipe section were presented.

The virtual commissioning allowed to verify that any part of the external surface of the tube could be accessed by means of movements of the robotic arm along the sides of the tube. It was demonstrated that it is possible to measure the external tube thickness anywhere on the external surface of the nuclear tube.

The improvements in the End-Effector design were significant and the results in the design were unprecedented with numerous innovations, among them: automatic application of the coupling liquid.

The technology was digitally validated in the laboratory as shown in figures 5, 6, 7. It is currently in the final phase of experimental validation and will be presented to the plant directors. In the future, if it is of interest, it may be applied within the plant itself.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are grateful for the financial support of Eletrobras Termonuclear – Eletronuclear, Amazônia Azul Tecnologias de Defesa, Fundação Parque de Alta Tecnologia da Região de Iperó, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES.

References

[1] USNRC-NUREG–2191. *"Generic Aging Lessons Learned for Subsequent License Renewal (GALL-SLR) Report"*, Final Report Vol. 2. US Nuclear Regulatory Commission, 2017.

[2] IAEA-SRS 82. Ageing management for nuclear power plants: International generic ageing lessons learned (IGALL), Rev. 1. *International Atomic Energy Agency, 2020.*

[3] S. A. Cancemi; R. L. Frano; Preliminary study of the effects of ageing on the long-term performance of NPP pipe. *Progress in Nuclear Energy*, 131, 103573, 2021.

[4] Q. Zhang; Y. Li; E. Lim; J. Sun. Real Time Object Detection in Digital Twin with Point-Cloud Perception for a Robotic Manufacturing Station. *Proceedings of the 27th International Conference on Automation & Computing*, University of the West of England, Bristol, UK, 1-3 September 2022.

[5] H. Zhou; S. Zhang; J. Zhang; C. Zhang; S. Wang; Y. Zhai; W. Li. Design, development, and field evaluation of a rubber tapping robot. *Journal of Field Robotics*, 39, 28–54, 2021.

[6] Tugal, H.; Cetin, K.; Petillot, Y.; Dunnigan, M.; Erden, M. S. Contact-based object inspection with mobile manipulators at near- optimal base locations. *Robotics and Autonomous Systems*, 161, 104345, 2023.

[7] J. Barbosa, P. Leitão, E. Adam and D. Trentesaux, "Dynamic self-organization in holonic multi-agent manufacturing systems: The ADACOR evolution," *Computers in industry*, v. 66, p. 99-111, 2015.

[8] J. Barata, L. Camarinha-Matos and M. Onori, "A multi-agent-based control approach for evolvable assembly systems,"

INDIN'05. 2005 3rd *IEEE International Conference on Industrial Informatics*, pp. 478-483, August 2005.

[9] E. Trunzer, A. Calà, P. Leitão, M. Gepp, J. Kinghorst, A. Lüder et. al., "System architectures for industrie 4.0 applications," *Production Engineering*, v. 13, n. 3, p. 247-257, 2019.

[10] Boschert, S.; Rosen, R. *Digital Twin—The Simulation Aspect*. In: Hehenberger, P.; Bradley, D. (Editors), Mechatronic Futures - Challenges and Solutions for Mechatronic Systems and their Designers. pag. 59-74, Spring, 2016.

[11] Bratchikov,S.; Abdullin, A.; Demidova, G. L.; Lukichev, D. V. Development of Digital Twin for Robotic Arm. 2021 *IEEE 19th International Power Electronics and Motion Control Conference (PEMC)* DOI: 10.1109/PEMC48073.2021.9432535, 2021.

[12] Nekoo, S. R.; Acosta, J. A.; Heredia, G.; Ollero, A. A benchmark mechatronics platform to assess the inspection around pipes with variable pitch quadrotor for industrial sites. *Mechatronics* 79, 102641, 2021.

[13] Lee, D. et al. Digital twin-driven deep reinforcement learning for adaptive task allocation in robotic construction. *Advanced Engineering Informatics*, v. 53, 1 ago. 2022.

[14] *Gartner Top 10 Strategic Technology Trends For 2018*. Disponível em: <https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018>. Acesso em: 20 fev . 2023.

[15] *What Is Digital Twin Technology - And Why Is It So Important?* Disponívelem: <https://www.forbes.com/sites/bernardmarr/2017/03/06/what-is-digital-twin-technology-and-why-is-it-so-important/?sh=1365c81c2e2a>. Acesso em: 20 fev. 2023.

[16] Tugal, H.; Cetin, K.; Petillot, Y.; Dunnigan, M.; Erden, M. S. Contact-based object inspection with mobile manipulators at near-optimal base locations. *Robotics and Autonomous Systems*, 161, 104345, 2023.

[17] Bratchikov,S.; Abdullin, A.; Demidova, G. L.; Lukichev, D. V. Development of Digital Twin for Robotic Arm. 2021 *IEEE 19th International Power Electronics and Motion Control Conference* (PEMC) DOI: 10.1109/PEMC48073.2021.9432535, 2021.

[18] Nekoo, S. R.; Acosta, J. A.; Heredia, G.; Ollero, A. A benchmark mechatronics platform to assess the inspection around pipes with variable pitch quadrotor for industrial sites. *Mechatronics* 79, 102641, 2021.

[19] Qiu, B. et al. A Feasible Method for Evaluating Energy Consumption of Industrial Robots. Proceedings of the 16th *IEEE Conference on Industrial Electronics and Applications, ICIEA* 2021. Anais...Institute of Electrical and Electronics Engineers Inc., 1 ago. 2021a.

[20] Lee, D. et al. Digital twin-driven deep reinforcement learning for adaptive task allocation in robotic construction. Advanced *Engineering Informatics*, v. 53, 1 ago. 2022.

[21] Partiksha; K., A. Robotic Tele-operation Performance Analysis via Digital Twin Simulations. 2022 *14th International Conference on COMmunication Systems and Networks, COMSNETS* 2022. Anais...Institute of Electrical and Electronics Engineers Inc., 2022.

[22] Fan, S. et al. A new approach to enhance the stiffness of heavy-load parallel robots by means of the component selection. *Robotics and Computer-Integrated Manufacturing*, v. 61, 1 fev. 2020.

[23] Huynh, H. N. et al. Modelling the dynamics of industrial robots for milling operations. *Robotics and Computer-Integrated Manufacturing*, v. 61, 1 fev. 2020.

[24] Vitalli, R. *Método de abordagem e estrutura geral do projeto de pesquisa desenvolvimento da RSTM.* Documento Número: UFABC- ROBOT-001-Rev00, 2023.

[25] Partiksha; K., A. Robotic Tele-operation Performance Analysis via Digital Twin Simulations. 2022 *14th International Conference on Communication Systems and Networks, COMSNETS* 2022. Anais...Institute of Electrical and Electronics Engineers Inc., 2022.

[26] Fan, S. et al. A new approach to enhance the stiffness of heavy-load parallel robots by means of the component selection. *Robotics and Computer-Integrated Manufacturing*, v. 61, 1 fev. 2020.

[27] Huynh, H. N. et al. Modelling the dynamics of industrial robots for milling operations. *Robotics and Computer-Integrated Manufacturing*,v. 61, 1 fev. 2020.

[28] SIEMENS PLM, *Process Simulate*. Germany. Siemens Product Lifecycle Management Software. 2018. Document ID: MT45215. Version: 15.1.2.

[29] ABB, *Product manual IRB 1600/1660*. Sweden. ABB Robotics. 06/01/2022. Document ID: 3HAC026660-001. Revision: AH.

[30] Wildgrube F, Perzylo A, Rickert M, et al. Semantic mates: Intuitive geometric constraints for efficient assembly specifications. 2023 *IEEE/ RSJ International Conference on Intelligent Robots and Systems (IROS)*; 2023.

[31] Rauen H. *Industrie 4.0 in practice* - Solutions for industrial applications. Frankfurt: VDMA Industrie 4.0 Forum; 2023.

[32] Karaman S, Frazzoli E. Sampling-based algorithms for optimal motion planning. *The International Journal of Robotics Research*.2024;30(7):846–894.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



ROGÉRIO VITALLI has PhD (2023) in Mechanical Engineering (Area: Mechatronics) from UNICAMP-FEM with a sandwich period in Israel and at the German Aerospace Center (Deutsches Zentrum für Luft und Raumfahrt, DLR, Robotik und Mechatronik Zentrum, Oberpfaffenhofen, Robotics and Mechatronics Center, RMC, Institute of Robotics and Mechatronics). He has experience in the area of Digital Twins, Advanced Manufacturing, Industry 4.0 and Mechatronics, working mainly on the following themes: Advanced Mechatronic Systems, Industrial Robotics, Modeling and Simulation of Robotic Manipulators, Forward Kinematics, Inverse Kinematics, Advanced Dynamics and Jacobians of Industrial Robots. He has published 14 articles in journals, 1 book, 4 book chapters, 6 articles in conference proceedings and 8 articles in newspapers. He has trained more than 200 roboticists, 9 specialists in robotics engineering. He is currently a Nuclear Engineering researcher at UFABC and a post-doctoral fellow at the University of São Paulo - USP.



JOÃO MOREIRA obtained a master's degree in nuclear science (1981) and a PhD in nuclear engineering (1984) from the University of Michigan, USA. He worked at the Institute for Energy and Nuclear Research (IPEN) between 1984 and 1989 and at the Navy Technological Center in São Paulo (CTMSP) between 1990 and 2006. In 2006 he moved to the academic field, joining the Federal University of ABC, where he is a Full Professor of Nuclear Energy. He has published 41 articles in journals, 3 books, 10 book chapters, 84 articles in conference proceedings and 4 articles in journals. He has trained 26 masters and 7 doctors and coordinated 2 research projects on the decommissioning of CNAAA (nuclear plants Angra 1, 2 and 3) for Eletronuclear.

Received: 22 September, 2024, Revised: 04 December, 2024, Accepted: 05 December, 2024, Online: 17 January, 2025

DOI: https://doi.org/10.55708/js0401002

Developing E-Methane Value Chain and Proper Greenhouse Gas (GHG) Accounting Rules Incentivizing Recycled Carbon Fuels

Ryota Kuzuki*, Mitsuhiro Kohara, Noboru Kizuki, Satoshi Yoshida, Yuta Nezasa, Yuki Tsuji

Division of Int'l Certification & Standards Harmonization, Planning dept., The Japan Gas Association

*Corresponding author: Ryota Kuzuki, Division of Int'l Certification & Standards Harmonization, The Japan Gas Association 1-15-12, Toranomon, Minato-ku, Tokyo, 105-0001, Japan, Contact No: +81-3-35020116, Email: kuzuki.ryota@gas.or.jp

ABSTRACT: Many countries and industries in the world are aiming to transition to a carbon neutral society. One of the means to achieve this is the use of recycled carbon fuels (RCFs) such as e-methane. Feasibility of an international value chain is being studied in a variety of regions. In particular, with regard to e-methane, demonstration projects are progressing in several regions around the world as well as feasibility studies of cross-border value chain of e-methane. Ideally, e-methane production sites should be located to secure abundant and inexpensive hydrogen and CO_2 as feedstocks, besides close to natural gas transmission pipelines, liquefaction facilities, and easily accessible to LNG exporting terminals. In order to stimulate investment in commercialization, development of internationally applicable GHG accounting rules is need at first. It is essential that the CO_2 emitted in combustion is recognized under international accounting rules as having environmental value attribute that does not increase CO_2 in the atmosphere. As one of the examples, the Japan Gas Association and related organizations have developed a certification scheme called "Clean Gas Certificate" and has applied it to several domestic e-methane production projects in Japan. For the transition period (2030-2050), the use of recycled CO_2 captured from factories and thermal power plants emissions are expected to be major feedstock. We propose GHG calculation and accounting method in consideration of avoiding double counting of recycled CO_2 in the entire supply chain. A calculation formula of the carbon footprint (CFP) of e-methane was proposed to one of the ISO's working groups and it was successfully included in the ISO 6338-1:2024. Next, it is necessary to harmonize the proposed methods with internationally common GHG accounting rules.

KEYWORDS: e-methane, methanation, carbon capture utilization, GHG accounting rules, GHG intensity, recycled carbon fuels, environmental attribute certificates, Clean Gas Certificate

1. Introduction

Many countries and industries in the world are aiming to transition to a carbon neutral society. One of the means to achieve this is the use of recycled carbon fuels (RCFs) such as e-methane. In [1], our previous paper " Promoting Introduction of E-methane for Carbon Neutral City Gas Supply in Japan" , focused on e-methane and overviewed about the gas industry's challenge of technological development and related initiatives such as net GHG emission calculation of e-methane. Following it, this paper describes the progress of the challenges of technological development, demonstrations, and feasibility studies for commercialization. The core concept is CO_2 capture and utilization (CCU) which enables carbon reduction. As an example, by use of e-methane, Figure 1 shows an image of the principle of RCFs. By replacing fossil fuels with RCFs, CO_2 emissions can be reduced.

In order to stimulate investment in commercialization, development of internationally applicable GHG calculation and accounting rules is essential. The originality of this paper is that it makes a proposal based on the authors' practices, while taking into account existing international standards to address this issue. This paper also introduces a certification scheme which has

developed to enable transferring the environmental value attributes of RCFs such as e-methane.

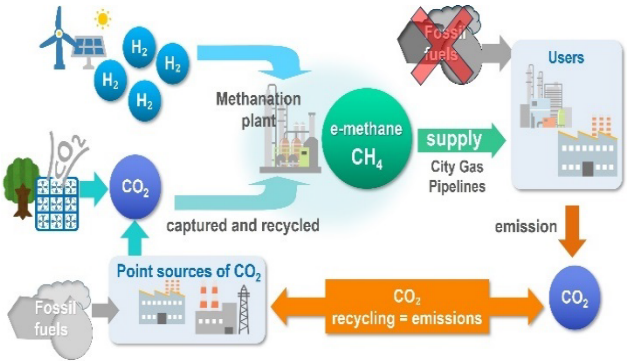


Figure 1: Image of methanation process for carbon neutral city gas supply

2. Expectations for e-methane and initiatives to date

In [2], the IPCC's 6th Assessment Report, released in April 2022, states "For greater compatibility with existing gas systems and appliances, the hydrogen can be combined with captured carbon dioxide to form methane and other synthetic fuels". In [3], a model analysis of IEA's "Net Zero by 2050" also estimates that 35% of global gas demand supplied by the gas grid in 2050 will be supplied by low-carbon gases, of which about 30% will be e-methane.

In September 2024, the IEA held a webinar focusing on e-methane for the first time, and in [4], Molnar presented survey looking over the global situation. Leading projects taken up by the IEA are shown in Figure 2. Among them, Tree Energy Solutions (TES) and Nordic Ren-Gas as well as the Japan Gas Association were invited as panelists and presented leading projects.

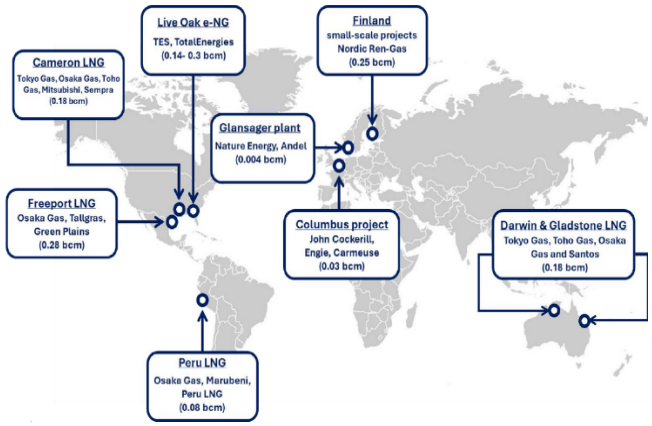


Figure 2: International partnerships drive e-methane projects (IEA's webinar on E-methane: a new gas for a net-zero future?)

In [5], Vercamen announced the formation of the e-NG Coalition, a framework for large companies to collaborate and advocate for their portfolio of e-methane projects in Japan, North America, Europe, and the Middle East.

In [6], Kujara, Nordic Ren Gas presented its plans of commercial production of e-methane at six sites in Finland to be operated from 2027, and announced that one of their project was selected by the EU Hydrogen Auction bid as the most competitive cost of hydrogen.

In [7], the Japan Gas Association presented about its Action Plan in June 2021, which sets an interim goal for the entire city gas industry to establish a value chain including overseas imports of e-methane covering 1% in city gas supply (approximately 400 million m^3) in 2030. And this includes a goal by 2050 of achieving carbon neutral city gas supply by e-methane as main feedstock (Figure 3).

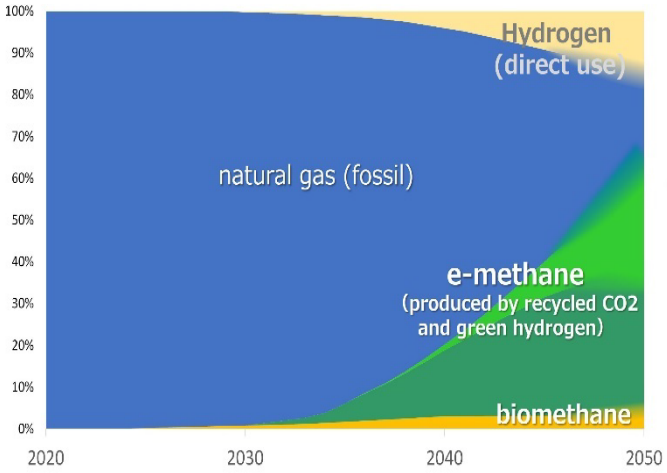


Figure 3: JGA's vision of transition of the share of feedstocks for Japan's city gas supply

In [8], European Biogas Association (EBA) published an Europe-wide study in September 2024. It reports a comprehensive survey on the e-methane projects in Europe. Figure 4 shows the increasing trend in number of e-methane production plants by country.

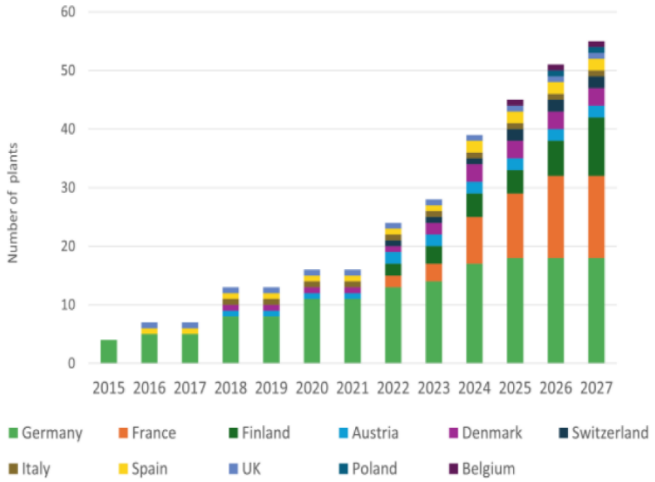


Figure 4: Growth in number of e-methane production plants in Europe (EBA; Mapping e-methane plants and technologies, 2024.9)

References above are appropriate to representing the latest situation. All of them suggest importance of proving the origin of the feedstock CO_2, but most of them are limited within biogenic CO_2, and there is few information of eligibility of fossil fuel-originated CO_2. Therefore, the attribution of emissions of fossil-originated CO_2 that is captured and used as feedstock are not discussed enough.

3. Harmonizing Internationally Applicable GHG Calculation and Accounting Rules for RCFs

3.1. Need for rules for cross-border transported RCFs

Ideally, the e-methane production site should have access to abundant and inexpensive hydrogen and CO_2 as feedstock and easy access to natural gas transmission pipelines, liquefaction, and LNG transportation. Several feasibility studies are currently underway in many countries and regions, including Nort America, Australia, Southeast Asia, Middle East, and South America. Figure 5 shows various feasibility studies (FSs) led by Japanese companies.

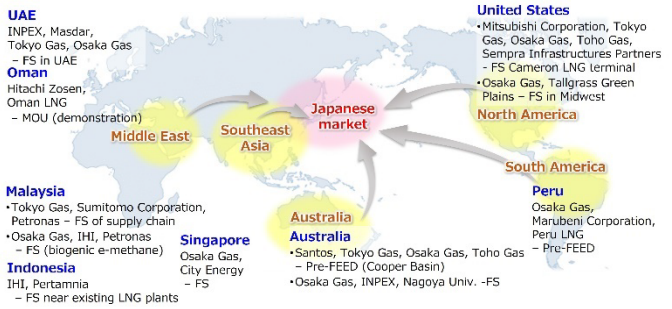


Figure 5: Various feasibility studies led by Japanese companies

As a result of site evaluations and feasibility studies conducted in various locations, they selected the area around the existing LNG export terminals as the most promising area because of their existing infrastructures for stable procurement of raw materials enabling early establishment of a value chain for e-methane. Some are expected for investment decisions. However, there is still no common international rule that can certify GHG calculation covering entire cross-border value chains and the transfer of environmental values of RCFs such as e-methane. Without them, there is concern that investment decisions would be delayed.

3.2. Existing international general rules for GHG calculation and accounting

During the transition period (from around 2030 to 2050), CO_2 captured from point sources is expected to become the major feedstock for producing e-methane. If the CO_2 emitted from point sources such as factories and power plants has already been accounted for and reported by the emitters, since the CO_2 would have been released into the atmosphere, when the e-methane is combusted, regardless of its origin (e.g. fossil or non-fossil), the emitted CO_2 should not be counted as additional emission in order to avoid double counting. This means that recycled carbon fuels (RCFs) such as e-methane have environmental value contributing CO_2 emission reduction, as stated in the 6th IPCC assessment report [2].

In order to promote social implementation of e-methane, common calculation method is needed to estimate the amount of CO_2 reductions by substituting fossil fuels. Internationally accepted rules for calculating GHG emissions include the IPCC Inventory Guidelines at the national level, the GHG Protocol Standards and Guidelines and ISO 14064, ISO 14067 (carbon footprint) at the corporate and product level. Method to calculate environmental value of RCFs such as e-methane should be harmonized with these existing rules.

One of the representing analysis of GHG emission of RCFs covering entire supply chain is by S. Morimoto et. al.. In [9], Morimoto calculated lifecycle GHG emission of e-methane imported from abroad to be 0.95-1.65 kg-CO_2/kg-CH_4 (nearly equals to 17.0-29.6 g-CO_2/MJ) under a condition that hydrogen is produced by renewable power.

Another example is that specific default values are being set for over 100 fuel types, including RCFs, which are being discussed at the International Maritime Organization (IMO). In [10], when the CO_2 is derived from the point sources such as flue gas after combustion of fossil fuels, calculation formulas are provided to subtract the amount of CO_2 emitted during combustion from the amount already reported by the original emitter. The calculation method accumulates the CO_2 emitted throughout the supply chain.

3.3. Development of calculation formula of GHG intensity in ISO

Based on the standards and guidelines mentioned above, we proposed a calculation formula to one of the ISO's Working Group. As shown in Figure 6, referring ISO 14067, the proposed formula is to reflect in carbon footprint of e-methane within the entire boundary of well to consumption (WtC) in consideration of the environmental value that it does not increase CO_2 in the atmosphere if the CO_2 feedstock meets certain criteria.

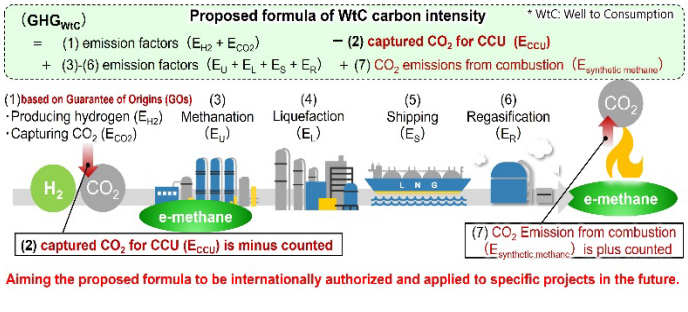


Figure 6: Proposed calculation formula of GHG intensity of e-methane

If the CO_2 feedstock is obtained by direct air capture (DAC) or biogenic origin, or if the CO_2 is already reported as GHG emissions before the e-methane production process, the captured CO_2 can be minus counted. This can offset the CO_2 emitted from combustion of the e-methane. In [11], this formula has been included in the ISO 6338-1:2024 (Calculations of greenhouse gas (GHG) emissions throughout the liquefied natural gas (LNG) chain Part 1: General) in January 2024.

This method was referenced when determining the upper limit of GHG emissions intensity of 49.3 g-CO_2/MJ, which is recognized as "low-carbon hydrogen, etc." as stipulated in [12] "Basic Act on a Hydrogen Society", it has been enacted by the Japanese government in May 2024.

3.4. Need for common rules to clarify the attribution of environmental value of RCFs

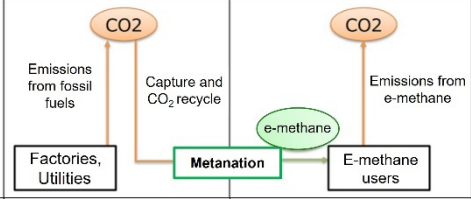
The next issue to be resolved is to develop common rule regarding who own the environmental value of e-methane, which does not increase atmospheric CO_2 even when combusted. In the case of biofuels, it can be commonly accepted that those who paid cost for biofuel own the environmental value. Recently, there is also a system in which the ownership of the environmental value is transferred by issuing certificates that separate the environmental value from the biofuels.

In the case of RCFs, unlike biofuels, the process of procuring CO_2 feedstocks and artificially synthesizing them is included in the value chain from production to consumption. The difference from biofuels and lack of common rules for RCFs causes uncertainty of the reward to the investment to developing a value chain for RCFs such as e-methane.

In Japan, discussions on how to handle CCU including e-methane are progressing. One of the most progressed discussions held in Japan was the Public-Private Council for Promoting Methanation led by the Ministry of Economy, Trade and Industry in 2021, and the Task Force for CO_2 Counting Rule under its auspices, have been held to discuss on this issue intensively.

In [13], the interim report issued by Ministry of Economy, Trade and Industry in 2022, as shown in Table 1, four cases, in which merits and demerits are included for the original emitters (captures' side) or the final users, or both, depending on account for emissions or not.

Table 1: Examples of CO_2 accounting among stakeholders by cases [13]

		Original emitter (capturers' side)	Users'-side
CO_2 behavior related to e-methane production and utilization			
CO_2 accounting for domestic production and consumption	Case 1	Account for emissions	No Emission
	Case 2	No emission	Account for emissions
	Case 3	Allocating emissions	Allocating emissions
	Case 4	No emission	No emissions

This report concluded that unless the environmental value is attributed to the users' side, e-methane will not be selected. Therefore, it was proposed that the users' side should account for as no emission. On the other hand, it suggested that complementary mechanism was necessary because there was a risk that production and supply of e-methane would not take place unless sufficient incentives were provided to those who capture CO_2 rather than emitting it to the atmosphere.

There is a progress to create a system in which the original emitters (capturers' side) can also claim environmental value for themselves, depending on the final use, by requiring proof of the captured CO_2 for CCU and proof of the final use.

Issues relating internationally applicable accounting rules has been raised in the intergovernmental frameworks. One of the examples is the Group of Seven (G7) Energy and Environment Ministerial Meeting held in Sapporo in April 2023. In [14], the communique stated that member countries would work together to resolve the issue (GHG accounting for RCFs).

As of now, the IPCC Inventory Guidelines nor the GHG Protocol, both are the most representative international rules, have not include any provisions regarding the treatment of CCU. As a result, even if RCFs is procured and used, users would not be rewarded since the CO_2 emitted during combustion still needs to be accounted for. As announced in [15] and [16], it is expected that discussions on CCU-related issues including RCFs will start from 2024 or 2025 and will be completed in a few years.

Meanwhile, as announced in [17], the EU's Renewable Energy Directive (RED) Delegated Act, which is a common rule for all member countries, came into effect in June 2023. It stipulates the calculation method for the environmental value of RCFs and RFNBOs (liquid or gaseous renewable fuels of non-biological origin). It would be possible that this will not only have an impact within EU, but also spread to the rest of the worlds.

4. Development of Environmental Attribute Certificates for e-methane

From around 2023, discussions have emerged on internationally applicable counting rules that clarify the attribution of environmental value for RCFs including e-methane and the direction is to contribute to motivate developing value chain for RCFs including e-methane.

In certain regions, certification schemes for RCFs are almost being finalized.

4.1. Voluntary Scheme recognized by European Commission

In [18], the European Commission has announced that several existing international voluntary schemes have completed the positive technical assessment (September 2024) to be recognized as institutions responsible for third-party review in line with the Delegated Act [17]. For example, in [19], ISCC (International Sustainable Carbon Certificate) describes requirements for fossil-derived CO_2 to be eligible as feedstock for RCFs and RFNBOs as follows; "Post-industrial (fossil) CO_2 captured from

industrial processes, which use fossil sources to deliberately produce electricity, heat, or materials (e.g., cement, iron and steel, petrochemical industry) must be verified that the CO_2 was not deliberately produced for use in the above-mentioned production processes."

4.2. Clean Gas Certificate (Japan)

(1) Overview of Clean Gas Certificate

We developed and launched "Clean Gas Certificate" scheme with gas-related organizations in Japan that certifies the attribution of the environmental value of e-methane and biogas, which is considered as not increasing CO_2 in the atmosphere even when combusted. As shown in Figure 7, this is environmental attribute certificate (EAC) making it possible to transfer the environmental value separated from e-methane or biogas and to bundle it with fossil natural gas.

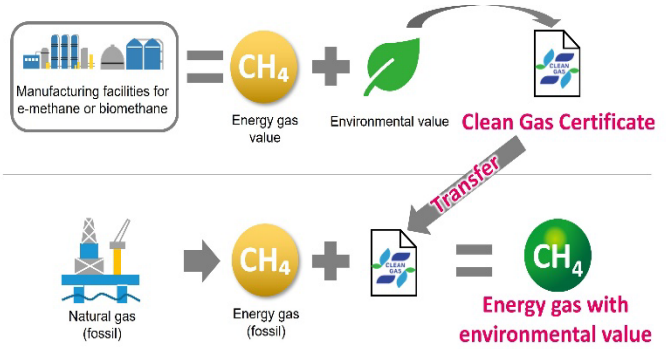


Figure 7: Image of function of "Clean Gas Certificate"

After a trial period in 2023, guidelines were developed, and the system was moved to actual operation as a private certification system from April 2024. In [20], details are posted on the website. With this scheme, e-methane and methane component in biogas are certified based on criteria. An sample image of the certificate to be issued is shown in Figure 8.

The attribute information included in the certificate are listed below. They have been decided based on surveys of some references of foreign certifying and registering systems for renewable fuels such as biogas.

- Serial number
- Issue date
- Production process
- Production period
- Name of the registered production facility
- Hydrogen suppliers' code
- CO_2 feedstock suppliers' code
- Name of the certificate issuer
- Type of clean gas (e-methane / biogas / mixed gas)
- Equivalent amount of clean gas (units: m^3 and MJ)
- Use of produced gas (injection into city gas pipelines / private consumption)



Figure 8: Sample image of "Clean Gas Certificate"

(2) Implementation of Clean Gas Certificate

In Japan, various efforts are underway to commercialize e-methane, including technology development, demonstration on both the supply and demand sides, with support from the Japanese government. Figure 9 shows demonstration projects of e-methane production by Japan's major city gas companies.

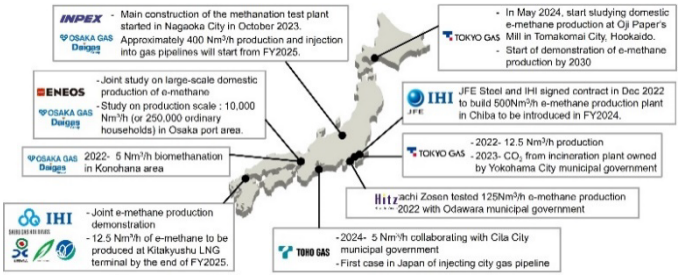


Figure 9: Demonstration projects of e-methane production in Japan

Some of the plants have already certified for Clean gas Certificates as below. Other projects planned in Japan are also expected to apply for the certification.

As announced in [21], in March 2022, Tokyo Gas Co., Ltd. began a methanation demonstration experiment at its research facility in Yokohama. Production capacity is 12.5Nm³/h. It procures electricity from renewable power sources and produce green hydrogen through water electrolysis. Having gained knowledge about the capabilities and issues of water electrolysis and the Sabatier methanation device, it has been working to improve the efficiency of the overall process. And in July 2023, it announced a new initiative to produce e-methane from CO_2 captured from the neighboring waste incineration plant owned by the Yokohama City municipal government, combined with the hydrogen obtained from its on-site water electrolysis device.

As announced in [22], in July 2022, Toho Gas has announced that it would work with the municipal government of Chita City in Aichi Prefecture to conduct a demonstration of methanation using CO_2 captured from biogas generated by sewage sludge treatment of the Chita City Southern Sewage Treatment Center and hydrogen produced by electricity generated from LNG cryogenic power generation. Production capacity is 5Nm³/h. The e-methane has been injected to the city gas supply network, which is for the first time in Japan.

As announced in [23], in December 2023, Saibu Gas announced that it started a demonstration project for methanation in Kitakyushu City in Fukuoka Prefecture using local raw materials in collaboration with nine other entities. Production is planned in FY2025 and with the capacity of 12.5Nm³/h. City gas customers, academia are

working together. Other city gas companies such as Hokkaido Gas, Hiroshima Gas, and Nihon Gas as well as the JGA also joined in this project aiming to transplant this model to other city gas supply areas.

As announced in [24], INPEX and Osaka Gas announced the start of a joint technology development project for the practical application of a methanation system. E-methane will be produced from FY2025 using CO_2 extracted from INPEX's gas field in Nagaoka City in Niigata Prefecture and will be injected into the company's existing gas pipeline. The e-methane production capacity is planned to be about 400 Nm^3/h.

As one of the representative examples of its users, as announced in [25], Osaka Gas has been assigned to supply Clean Gas Certificate bundled city gas to the facilities located in the EXPO 2025 Osaka, Kansai during the period of the event.

4.3. Aiming for broader internationally applicable schemes

Broader than above cases, it is necessary to operate the scheme based on internationally applicable CCU-related accounting rules to be developed in the near future. it is necessary to built in on the existing LNG supply chain that include tracking information on the guarantee of origins (GOs) of the feedstocks, CO_2 and H_2, and the GHG intensity of e-methane shown in Figure 6. Figure 10 shows an image of e-methane certificates in circulation.

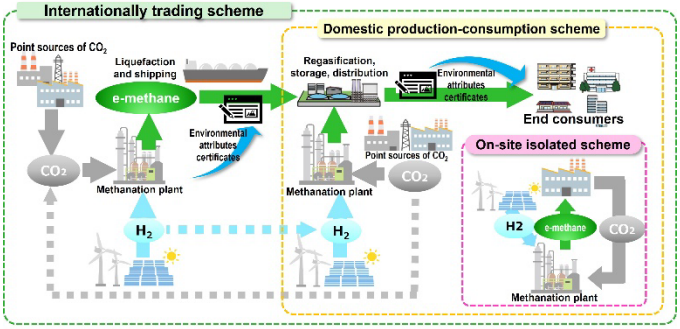


Figure 10: Image of environmental value transfer schemes in the existing LNG value chain including environmental attributes certificates

There are some notable developments regarding the IPCC Inventory Guidelines and GHG Protocol, which are the most commonly referenced GHG calculation and accounting rules in the world.

The IPCC has announced that it will hold expert meetings focusing on CO_2 removal and CCUS in the IPCC's Inventory Task Force through 2027, and will revise the inventory guidelines. As announced in [26], At the expert meeting in July 2024, "Address potential double-counting from the use of synthetic fuels (e.g. efuels)double counting" was raised as one of the issues to be considered.

The GHG Protocol is due for a major revision by the second half of 2026. In order to propose the addition of new rules on CCU to the standards and guidance, in [27], we submitted a proposal (open letter) to the secretariat in January 2024, signed by 18 voluntary organizations. As published in [28], we also appealed for the need for rules related to CCU at a session attended by the representative of the GHG Protocol Secretariat at the annual general meeting hosted by a European organization that promotes CCU, and he responded positively to our appeal.

Since diverse stakeholders intervene, a platform is needed to manage the history of GHG emissions tracking, verification, and transfer of the value of reduction credits at stages along the value chain. For example, as announced in [29], Osaka Gas Co., Ltd., Mitsubishi Heavy Industries, Ltd. and IBM Japan, Ltd. are jointly conducting a proof-of-concept (PoC) of a system to visualize CO_2 emissions throughout the supply chain and the environmental value of e-methane (Figure 11). It is expected that such fundamental technology will be developed at an early stage to serve as a platform for tracking, verifying, and managing the transfer of environmental values.

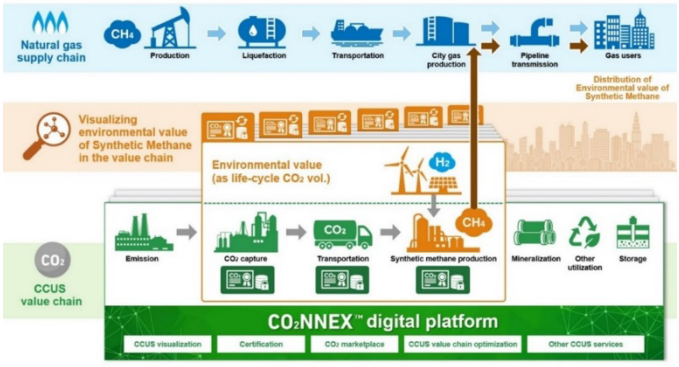


Figure 11: Example of digital platform tracking GHGs in e-methane supply chain

We will continue to monitor future discussions and work to harmonize our schemes with their revisions.

5. Conclusions

This paper focuses on e-methane, which is a leading RCFs produced from hydrogen derived from renewable energy and recycled CO_2, and which is attracting particular attention in Japan, Europe and North America because it can make the most of the existing LNG value chain. By leveraging the advantages of drop-in fuels, it is possible to reduce social costs and contribute to a smooth transition to a carbon-neutral society.

In this paper, we strongly advocate the need for international common rules for GHG calculation and accounting for RCFs through cross-border supply chains, which is a common issue for the social implementation of RCFs. In the early stages, the value chain of RCFs, including e-methane, is expected to grow within regional economic blocs such as Japan, EU and APEC, and in specific industries such as international shipping. As a result of certification of GHG intensity and in the carbon pricing system, market prices will be formed for RCFs. Based on the above circumstance, as a first step, we have

standardized the GHG intensity calculation formula for e-methane at the ISO and have also started to operate the Clean Gas Certificate scheme, which certifies the environmental value of e-methane produced at demonstration plants in Japan. Through these initiatives, we believe that the knowledge of trade-offs between GHG intensity and cost will contribute to form global RCFs market.

We especially expect academic world to scientifically clarify trade-off relationships that are more versatile. By sharing this knowledge, industrial sector is expected to work towards the formation of a global RCFs value chain. Everyone should collaborate to establish appropriate rules for a smooth transition to a carbon neutral society.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] R. Kuzuki, et.al., "Promoting Introduction of E-methane for Carbon Neutral City Gas Supply in Japan", *11th International Conference on Power Science and Engineering (ICPSE)*, 2022, doi: 10.1109/ICPSE56329.2022.9935370

[2] IPCC, "The Sixth Assessment Report of the IPCC, Working Group III Report, Climate Change 2022 Mitigation of Climate Change, pp.6-57, 2022 https://www.ipcc.ch/report/ar6/wg3/downloads/report/IPCC_AR6_WGIII_FullReport.pdf

[3] IEA, "Net Zero by 2050 A Roadmap for the Global Energy Sector", 2021, https://iea.blob.core.windows.net/assets/deebef5d-0c34-4539-9d0c-10b13d840027/NetZeroby2050ARoadmapfortheGlobalEnergySector_CORR.pdf

[4] G. Molnar, "E-methane: a new gas for a net-zero future?", *Webinar on IEA Low-emissions Gases Work Programme, 2024*, https://iea.blob.core.windows.net/assets/9b86ac2a-2055-4eac-9f93-6ab379554d6d/IEA_E-methaneanewgasforanet-zerofuture.pdf

[5] Y. Vercamen, "eNG's role for net-zero future", *Webinar on IEA Low-emissions Gases Work Programme, 2024*, https://iea.blob.core.windows.net/assets/9b86ac2a-2055-4eac-9f93-6ab379554d6d/TESIEAe-methanewebinar05Sep2024.pdf

[6] S. Kujara, "E-methane market prospects in Northern Europe", *Webinar on IEA Low-emissions Gases Work Programme, 2024*, https://iea.blob.core.windows.net/assets/9b86ac2a-2055-4eac-9f93-6ab379554d6d/Ren_Gas_E-methane_presentation_IEA_2024_0902.pdf

[7] The Japan Gas Association, Carbon Neutral Challenge 2050 (Press Release), 2020, https://www.gas.or.jp/pdf/en/newsrelease/newsrelease_20211101.pdf

[8] Europan Biogas Association, "Mapping e-methane plants and technologies – The role of e-methane it the total energy mix", 2024, https://www.europeanbiogas.eu/wp-content/uploads/2024/09/Mapping-e-methane-plants-and-technologies.pdf

[9] S. Morimoto et.al., "Scenario assessment of implementing methanation considering economic feasibility and regional characteristics", 2022, *Journal of CO_2 Utilization*, vol.58, 101935, doi.org/10.1016/j.jcou.2022.101935

[10] R. Miyake, "Overview of IMO's guideline of marine fuel life cycle GHG intensity (in Japanese)", pp.26-28, *2024 Class NK Technical Seminar*, 2024, https://www.classnk.or.jp/hp/pdf/research/seminar/2024/seminar_005.pdf

[11] ISO, "ISO 6338-1:2024 Calculations of greenhouse gas (GHG) emissions throughout the liquefied natural gas (LNG) chain Part 1: General", 2024, https://www.iso.org/standard/87114.html

[12] CCR (carbon capture and reuse) Study Group; "A Study on the Carbon Intensity of Synthetic Methane (e-methane)", *12th Public-Private Council for Promoting Methanation* (only Japansese), 2024, https://www.meti.go.jp/shingikai/energy_environment/methanation_suishin/pdf/012_05_01.pdf

[13] Interim report on CO_2 emission accounting in combustion of synthetic methane, Task Force for CO2 Counting Rules of the Public-Private Council for Promoting Methanation (only Japansese) 2022, https://www.meti.go.jp/shingikai/energy_environment/methanation_suishin/co2_tf/pdf/20220322_1.pdf

[14] G7 (Group of Seven) Climate, Energy and Environment Ministers' Communiqué Sapporo, 2023, 27, https://www.meti.go.jp/press/2023/04/20230417004/20230417004-1.pdf

[15] IPCC, Decisions adopted by the Panel, *60th Session of the IPCC*, 2024, 9, https://www.ipcc.ch/site/assets/uploads/2024/02/IPCC-60_decisions_adopted_by_the_Panel.pdf

[16] GHG Protocol, GHG Protocol Corporate Suite of Standards and Guidance Update Process, https://ghgprotocol.org/ghg-protocol-corporate-suite-standards-and-guidance-update-process

[17] European Union, ANNEX Methodology for determining greenhouse gas emissions savings from renewable liquid and gaseous transport fuels of non-biological origin and from recycled carbon fuels, Official Journal of the European Union, Volume 66, (June 20 2023) 23-33

[18] European Commision Website, "Voluntary schemes set standards for the production of sustainable fuels and gases" https://energy.ec.europa.eu/topics/renewable energy/bioenergy/voluntary-schemes_en

[19] ISCC, ISCC PLUS Ver 3.4.2, 2024, 13-15 https://www.iscc-system.org/wp-content/uploads/2024/03/ISCC-PLUS_v3.4.2.pdf

[20] Cleasn Gas Certificate Committee Website (only in Japanese) https://www.clean-gas-certificate.com/

[21] Tokyo Gas Co., Ltd., Towards a carbon-neutral, decarbonized society in 2050 - methanation demonstration tests begin (Press Release (in Japanese)) , 2022, https://www.tokyo-gas.co.jp/news/press/20220323-02.html

[22] Toho Gas Co.,Ltd., Methanation demonstration test using CO_2 from Biogas in collaboration with the minicipal government of Chita City (Press Release (in Japanese)), 2022, https://www.tohogas.co.jp/corporate-n/press/1229823_1342.html

[23] Saibu Gas CO., Ltd., etc.; Start of a methanation demonstration project using local materials (Press Release (in Japanese)) ,2023, https://hd.saibugas.co.jp/news_release/detail/2023/nr040.html

[24] INPEX Co.,Ltd and Osaka Gas Co.,Ltd, INPEX, Osaka Gas to Commence Technical Development Business on CO2 Emission Reduction and Practical Application of Effective CO2 Use Through One of the World's Largest Methanation Operations (Joint Press Release), 2021, https://www.osakagas.co.jp/company/press/pr2021/1300478_46443.html

[25] Osaka Gas Co., Ltd., Contribution to the carbon neutrality of the 2025 Japan International Exposition through the first Clean Gas Certificate (Press Release (in Japanese)), 2024, https://www.osakagas.co.jp/company/press/pr2024/__icsFiles/afieldfile/2024/10/31/241101_1.pdf

[26] IPCC, "IPCC Expert Meeting Report on Carbon Dioxide Removal Technologies and Carbon Capture, Utilization and Storage", 2024, https://www.ipcc-nggip.iges.or.jp/public/mtdocs/pdfiles/2407_CDR_CCUS_Report.pdf

[27] The Japan Gas Association and 17 organizations and companies, "Proposal for Inclusion of Market-Based Accounting in Scope 1 of the GHG Protocol Update" (open letter to the GHG Protocol Secretariat) ,2024, https://www.gas.or.jp/pdf/en/newsrelease/newsrelease_20240131.pdf

[28] GHG Protocol Website "Event Recording: Reporting Emissions in Carbon Capture and Utilization (CCU) Applications", 2024, https://www.youtube.com/watch?v=usB47sH-320

[29] Osaka Gas Co., Ltd., Mitsubishi Heavy Industries, Ltd., IBM Japan, Ltd., Three Partners to Jointly Undertake Proof of Concept (PoC) Applying "CO2NNEX™", a Digital Platform for CO_2 Accounting across its Supply Chain for Synthetic Methanes (Joint Press Release), 2022, https://www.mhi.com/news/221021.html

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/bysa/4.0/).



Ryota Kuzuki has done his master's degree of precision engineering from the University of Tokyo in 1989. He has received Ph.D. degree in engineering from graduate school of Keio University in 2011.

He joined Tokyo Gas Co., Ltd. in 1989 and worked mainly for gas supply safety system design, corporate planning, and commercial and industrial sector's gas marketing. He currently serves as Division Head, Int'l Certification & Standards Harmonization, JGA



Mitsuhiro Kohara has done his master's degree in mechanical engineering from Graduate School of Engineering, Osaka University in 2001.

He joined Osaka Gas Co Ltd, where he was mainly involved in the design and construction of liquefied natural gas receiving terminals and was engaged in technical development. He currently serves as Manager of the Int'l Certification & Standards Harmonization Group, Planning Dept, JGA



Noboru Kizuki has done his master's degree in machinery engineering from National University in 2000.

He joined Osaka Gas Co., Ltd. in 2000, as an engineer of measuring and control system design and operation for LNG plants (2000-2010), Strategic Planning section (2010-2021). He currently serves as a Manager of Gas Production Group, Engineering Dept. and Int'l Certification & Standards Harmonization, Planning Dept, JGA



Satoshi Yoshida has done his bachelor's degree of architecture from Waseda University. He has done his master's degree of urban planning from University of Washington.

He is working for Japan's city gas industry since 1982 and currently serve as Senior Adviser to JGA.

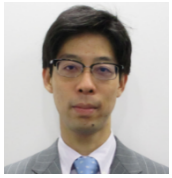
He served as Senior Council Member for "Green Purchase Network", Japan (2007-2012), Chair, Sustainable Committee, International Gas Union (2012-2015).

He is an Executive Committee Member at International Gas Union since 2019. He is also a Senior Certified Architect registered in Tokyo.



Yuta Nezasa has done his master's degree in civil engineering from Graduate School of Engineering, Univ. of Nagoya, 2014.

He joined Toho Gas Co., Ltd., from 2014, he worked mainly for gas supply pipeline construction, disaster prevention planning, and corporate planning. He currently serves as Chief of Int'l Certification & Standards Harmonization, Planning Dept., JGA.



Yuki Tsuji has received his diploma from the School of Law at Keio University in 2014.

He joined the Japan Gas Association from 2014, and mainly worked for general affairs and residential sector's gas marketing.

He currently serves as Chief of Int'l Certification & Standards Harmonization, Planning Dept., JGA

Received: 22 September, 2024, Revised: 01 December, 2024, Accepted: 02 December, 2024, Online: 17 January, 2025

DOI: https://doi.org/10.55708/js0401003

How to Analyze the Introduction Strategies for Radically New Technological Innovations? The Case of Autonomous Shipping

Mariah Kurtinaitis Joukes*, Roland Ortt, Mark de Bruijne, Linda M. Kamp

Faculty of Technology, Policy & Management, TU Delft, Delft, 2628 BX, The Netherlands

E-mail: mariahkurtinaitis@gmail.com (M. K. Joukes), J.R.Ortt@tudelft.nl (R. Ortt), M.L.C.deBruijne@tudelft.nl (M. d. Bruijne), L.M.Kamp@tudelft.nl (L. M. Kamp)

* Corresponding author: Mariah Kurtinaitis Joukes, E-mail mariahkurtinaitis@gmail.com

ABSTRACT: This article explores Maritime Autonomous Surface Ships (MASS) through the Technological Innovation Systems (TIS) framework to identify strategies for creating a market for this emerging technology. It examines the current state of TIS components for MASS and the factors shaping their development. Through a literature review and expert interviews, the article highlights key barriers, particularly the cost-benefit ratio, and suggests niche strategies to address barriers. Proposed strategies include the top niche strategy, lead user strategy, and hybridization niche strategy, aimed at overcoming challenges and facilitating broader market adoption of MASS in the future.

KEYWORDS: Autonomous ships, MASS, Technological Innovation System, niche strategy

1. Introduction

Radically new technological innovations represent an advance in technology that is so significant that either price/performance ratios are altered dramatically, or entirely new kinds of applications are made possible [1], [2]. Historical examples of such innovations in the automotive industry are hydrogen cars compared to cars with internal combustion engines. The diffusion of radical technological innovations differs from incremental innovations [3]. Radical innovations seldom start diffusing smoothly in an existing market. After the first introduction, an erratic process of introducing and withdrawing specific versions of the innovation in niche markets can be witnessed. It is remarkable how long it takes before radically new technological innovations start diffusing on a large scale [4-6]. This erratic process is often caused by a market that is full of barriers to large-scale diffusion. Another difference between radical and incremental innovations is the change in the socio-technical system required for large-scale diffusion. Radically new technological innovations often entail new complementary products and services, new customer groups and adapted institutional arrangements, for example. This means that a new market arrangement must emerge before large-scale diffusion of that radically new technological innovation is possible. The ultimate market arrangement is often unknown early on, because of uncertain progress in the development of technologies, complex and unpredictable patterns of competition between old and other new technologies. The socio-technical system changes around a radically new technological innovation can be explored using a Technological Innovation System framework [7,8].

This article focuses on Maritime Autonomous Surface Ships (MASS) in order to illustrate how to analyze the wider system around an upcoming market and derive introduction strategies into that market. It takes the work of two previous articles as a starting point. In [9], we concluded that the diffusion of MASS is hampered by several barriers and MASS is thus applied in small scale niche applications only, such as survey and ferry transport. So, MASS has not yet started to diffuse on a large scale [9]. In [10], we analyzed the wider sociotechnical system around the upcoming market for MASS. We used the

Technological Innovation System (TIS) framework that can be used by maritime managers and researchers. This framework combines technological, economic, institutional and social aspects that are important for large-scale diffusion and relevant for strategy formulation. It contains seven so-called TIS building blocks of the socio-technical system from which new markets emerge [11]. We concluded that some important building blocks are still fully or partially absent that are required for large-scale diffusion of MASS [10].

However, a knowledge gap remains: to date, the literature has not explored the strategies that can be employed to spur the diffusion of MASS. The current article will investigate that. It extends the previous work by showing how several underlying conditions called influencing factors influence the status of the TIS building blocks for MASS. Based on the insight into the status of the TIS building blocks and the influencing factors we derive strategies to spur the diffusion of MASS. We focus on the application of MASS in commercial shipping rather than pleasure craft and on the societal and technological aspects that enable MASS as a whole, without detailing the enabling technologies that compose the various types of autonomous vessels. This leads to two research questions:

RQ1: What is the status of the building blocks that together form a Technological Innovation System (TIS) around MASS and how is that status influenced by influencing factors?

RQ2: How can potential (niche introduction) strategies be derived from the analysis of the TIS around MASS?

The research questions are scientifically relevant because the TIS framework provides a framework that can be used to analyze the context at the time a new market is emerging or does not even exist yet. That means the framework is an addition to the extant market research literature. The framework represents a (socio-technical) system perspective, but instead of most of the innovation system literature, it is not used to formulate policies for governments but to formulate strategies for companies [11]. The research questions are managerially relevant because the TIS framework and the strategies derived thereof provide an analytical tool for a situation in which standard market analysis instruments fail, and standard strategies are not applicable. As a result, many companies are found to fail in the situation of new and emerging markets for radically new technological innovations.

To answer these questions, we used a combination of literature research and two rounds of interviews with MASS researchers and practitioners. The unstructured and semi-structured interviews gave a broad perspective of the barriers and a detailed view of the barriers and their underlying influencing factors. To answer research question 2, we used the approach outlined in the work of [11].

Our article is organized as follows: the theoretical notions are described in section 2, the method in section 3 after which results are presented in section 4. Conclusions, discussion and future research avenues are in section 5.

2. Theoretical background

In this section, we present a general TIS framework following the work of [11]. The framework is used to assess potential barriers and their influencing factors during the process of development and diffusion of radically new technological innovations, such as MASS. In practice, barriers to large-scale diffusion lead to a pattern of development and diffusion that differs from the well-known diffusion pattern introduced by [12]. Instead of a smooth S-shaped diffusion curve, a more erratic pattern can be witnessed for diffusion processes of radically new technologies [13]. After that, we explain how the framework can be used to derive possible strategies during an upcoming and turbulent market around a radically new technological innovation. In general, the notion of a TIS framework has been criticized in several ways [14]. One of the limitations revolves around the lack of attention for company strategies. Innovation system frameworks traditionally focus on governments and are mostly meant to inform policy formation rather than strategy formation. We tackle this limitation by creating a TIS framework from a company perspective that is shaped to facilitate strategy formulation.

2.1. A general TIS framework to formulate strategies

The framework is built up in a few steps and is composed of seven building blocks and seven influencing factors [11]. Each of these building blocks is essential for the emergence of a complete market and for the large-scale diffusion of a technological innovation in that market. Conversely, in the case of absence, incompleteness or incompatibility of one or more of these building blocks, large-scale diffusion is seriously hampered. The building blocks are:

1. Product performance and quality;
2. Product price;
3. Production system;
4. Complementary products and services;
5. Network formation and coordination;
6. Customers;
7. Innovation-specific institutions.

The explanation of each of the building blocks is summarized in Appendix 1.

These building blocks are important to formulate strategies. If all building blocks are complete and compatible, then large-scale diffusion is possible. In that situation, companies may consider adopting large-scale production and marketing strategies to sell the innovation. In contrast, if none of the building blocks is in place, complete and compatible, then introduction strategies are not a realistic option. If some building blocks are incomplete or incompatible, then these building blocks represent barriers to large-scale diffusion. In such cases, companies can sometimes adopt specific niche strategies that circumvent the barriers. This means that analysis of the building blocks is important to decide about specific aspects of introduction strategies: the timing of introduction (early versus late) and the scale of introduction (niche versus large-scale introduction).

Timing and scale of introduction can be derived from the status of the building blocks and are important aspects of an introduction strategy, but do not suffice. Especially in the face of an emerging and incomplete market it is important to deliberately formulate a niche strategy that fits the situation. That means we need more information to formulate the type of niche introduction strategy. To do so it is important to explore the causes of the incomplete or incompatible building blocks. These causes, referred to as influencing factors, are:

1. Knowledge and awareness of technology;
2. Knowledge and awareness of application and market;
3. Natural, human and financial resources;
4. Competition;
5. Macro-economic and strategic aspects;
6. Socio-cultural aspects;
7. Accidents and events.

The explanation of each of the influencing factors is summarized in Appendix 2.

If combinations of the influencing factors and building blocks are systematically explored several niche introduction strategies come to the fore. All these niche introduction strategies can be found in practice, as can be illustrated by historical cases of technological innovations. Different types of such niche introduction strategies are outlined in Table 1.

Table 1: General niche introduction strategies [15,16]

Niche Strategies	Definition
Demo Experiment and Develop Niche Strategy	Demonstrate the product in a controlled environment to enhance public knowledge about it and use the experiments to further develop the product quality. Relevant when knowledge of the technology is lacking.
Top Niche Strategy	Offer a limited number of products to the specific top end of the market that is not price-oriented until production capacity is feasible. Appropriate when knowledge of the technology is lacking and it affects its price.
Subsidized Niche Strategy	Subsidize the product for a particular segment if its use is considered socially important, but it is too expensive for the target customer. Applicable when the product is too expensive for the market, or when resources are scarce.
Redesign Niche Strategy	Introduce a simpler version of the product that can be produced with limited resources and sold at a lower price to explore the market. Appropriate when the product is too expensive to the market, when resources are scarce, or to fit socio-cultural aspects of suppliers and customers.
Dedicated System Niche Strategy	Offer the option of using the high-tech product independently from complementary products. Suitable when lack of knowledge about the technology affects the availability of complementary services.
Hybridization or Adaptor Niche Strategy	Combine the new product with the old one, allowing the use of all complementary products and services, or make the new product compatible with the existing complementary services. Applicable when resources are lacking, affecting the availability of complementary products.
Educate Niche Strategy	Educate customers and suppliers about the high-tech product to increase knowledge and awareness of the product capabilities. Useful when knowledge of the technology is lacking, affecting availability of suppliers and customers.
Geographic Niche Strategy	Introduce the high-tech product where laws and rules are less strict, or where customers, suppliers and

Niche Strategies	Definition
	resources are already available. Suitable when socio-cultural or macro-economic aspects affect the customers, and when institutional aspects pose a barrier to technology adoption.
Lead User Niche Strategy	Partner with lead users and innovators to co-develop products that innovators are willing to experiment with. Applicable when knowledge of the application is lacking, affecting a clear view of customer's applications.
Explore Multiple Markets Niche Strategy	Introduce the high-tech product in different customer applications to stimulate the explorative use in new functions. Relevant when knowledge of the application is missing, reducing the visibility of product application, usage pattern and benefits.

3. Methods

Given the lack of specific literature about the innovation management aspect of autonomous shipping and the complexity of the maritime industry, we deemed exploratory research the most appropriate method for this article. So, we based our analysis mainly on qualitative data in the form of literature review and semi-structured interviews. However, to safeguard a broad empirical basis using these two qualitative approaches, we interviewed a wide variety of experts and practitioners, and we interviewed them in two rounds. In the first round of interviews, open questions were posed regarding relevant actors and factors and possible barriers to large-scale use of MASS. In the second round, after consolidating the list of actors and factors, we returned to the experts to have their verdict regarding the barriers among those actors and factors for the purpose of validation.

The present article was based on the same research method as [10]. Therefore, the method section of the present article is very similar to the method section in [10].

We started our research with a literature review. This was necessary to understand the context in which autonomous shipping is being developed, as well as the barriers to MASS technology development already identified in the literature. In this way, we came to understand the factors that prevent or encourage MASS large-scale adoption by the maritime industry. This literature review also allowed us to adapt the general list of barriers to the specific context of MASS. For this literature review we made use of textbooks, conference proceedings, white papers, and scientific articles available in the TU Delft Library database and on the internet. We used the keywords "Maritime", "Ships", "Autonomous"," Innovation", "Diffusion" and other similar terms in our search.

Because of the exploratory nature of this research, the broad variety of MASS usage in the maritime industry, the novelty of MASS technology and the lack of data regarding the barriers to its adoption by the maritime industry, next to the literature review we also made use of semi-structured interviews. The goal of semi-structured interviews is to have a specific set of questions for all interviewees but still allow room for further questions within the topic. This allows the interviewer to adapt the question list to fit the expertise of the particular interviewee [17]. We created a list of interview questions aimed at clarifying aspects of the literature available and acquiring more in-depth data about MASS and the TIS Framework actors and factors for the case of MASS.

3.1. Selection of interviewees

We performed 14 semi-structured interviews with MASS researchers, sailors, employees of companies that develop MASS technology, a ship owner, a Navy System Integrator involved in MASS projects and a shipyard manager. We chose these professionals based on their experience in MASS or their understanding of the maritime industry characteristics as shown in their LinkedIn profile, scientific articles or white papers written about the research topic. We contacted the interviewees using social media, or via the website of the companies for which they work. We asked each interviewee to refer to possible other interviewees, creating a snowball effect. The reason for choosing the only one of our interviewees not connected to the merchant marine sector, the Naval System Integrator, was not only his broad MASS knowledge acquired in the MASS joint industry project but also because the navy is known as the testbed for maritime technology, acting as an innovator according to the diffusion scale defined by [18]. We made sure to have a broad variety of interviewees in terms of their background and expertise. However, all our interviewees were based in the Netherlands. That can be a shortcoming but the effect of this is probably limited because the maritime industry in the Netherlands is among the industrial clusters that is closely involved with MASS and all our interviewees had an international perspective.

3.2. Data Analysis

We recorded all interviews, except when the interviewee preferred not to be recorded. In this specific case, we used our interview notes as material for further analysis. We created an automated transcript of the recorded interviews by making use of the program Otter AI, a program that transforms voice into text. After that, we reviewed the transcripts and interview notes according

to the content and we assigned codes to different categories for further data analysis. A complete list of keywords and databases used in the literature review of this article, the list with all interview questions, an example summary of the interviews, and the table with codes and categories of the interview analysis are available upon request.

Both the data from the literature and data from the interviews are mainly qualitative and hence our analysis is also qualitative. We started with a generalizable TIS framework. We used a combination of qualitative data and well-informed opinions (also qualitative in nature) voiced by experts to adapt that framework to the specific maritime industrial context. After adapting the framework, we could qualitatively assess the main barriers. For some of the barriers, more detailed and quantitative data could have been gathered but that would not have resulted in a better comparison of the relative importance of these barriers. Neither would it form a better basis to select relevant niche introduction strategies.

4. Results

As presented in the Theory section, many breakthrough technologies require the presence of specific actors and factors, the so-called TIS building blocks. This section presents and explains the (f)actors affecting MASS specifically, based on our desk research and interviews.

4.1. RQ1: What is the status of the building blocks that together form a Technological Innovation System (TIS) around MASS and how is that status influenced by influencing factors?

4.1.1. Quality & Performance

Quality and performance are inherent to each product. For MASS, its quality is defined by its capacity to increase ships' performance in terms of reliability, efficiency, and predictability. These performance measures were mainly rated by our interviewees as very important. However, MASS has not yet proven its performance in all these aspects, rendering this building block partially present [10].

4.1.2. Cost-Benefit Ratio

The cost-benefit ratio for MASS is calculated in terms of return on investments (ROI), which interviewees believe can be achieved by crew reduction and/or and increased space for cargo because of MASS implementation onboard. Crew reduction is not (yet) regulated, making the business case for MASS adoption unclear to ship owners. Technology providers, on the other hand, ensure there is a valid return on investment to MASS, regardless of crew reduction. All interviewees considered cost-benefit ratio an extremely important building block. The Royal Dutch Navy interviewee is the only exception since the focus of the navy lies on reducing the number of crew members and less on costs. It is valid to note that navy vessels are not obligated to follow merchant marine regulations and therefore can determine their crew size based on their needs and not on legal requirements. Most interviewees agree that a beneficial cost-benefit ratio is still missing in the MASS TIS [10].

4.1.3. Production System

MASS is not expected to ever reach mass production, since ships are not built in the same numbers as cars. Our interviewees consider that a MASS production system is the capability to produce standardized items and have a set of hardware and software standards, regulated by accredited institutions. These standards do not exist yet, but the equipment is currently being built and sold. Therefore, this building block was considered present and moderately important [10], since the implementation of standards in the form of technology battles is common within technology development [19].

4.1.4. Network Formation & Coordination

When customers evaluate the value of a technology, they consider the perceived value of having a network that can or will increase the product's value in the future [20]. Our interviewees recognize there is little collaboration among companies offering MASS and the large ones create a closed loop of products, which shields smaller companies from a possible network. Knowledge is also not shared among companies offering MASS. Despite not being considered important, 'network formation and coordination' was recognized as a building block to MASS diffusion and deemed partially present [10].

4.1.5. Customers

The customers for MASS technology are ship owners willing to purchase and install the technology in their vessels. Without customers, there is no technology diffusion. Companies offering MASS claim to have sold and installed equipment; other companies mentioned potential customers and current clients interested in the technology. However, there is no widespread interest in the technology (yet). Hence, the building block is extremely important and only partially present [10].

4.1.6. Innovation Specific Institutions

MASS is not yet regulated by the International Maritime Organization (IMO), which means its conventions that regulate equipment and manning onboard vessels worldwide also do not address MASS.

Most interviewees see this lack of regulation as a barrier to MASS diffusion. Given the strong weight of the IMO recognition of the technology and its impact on influencing factors described further in this section, this building block was evaluated as extremely important.

4.1.7. *Safety*

One interviewee described safety as a preconditional item that can only be increased onboard. From his perspective, every change on a vessel must increase safety, otherwise, it is not acceptable. All other interviewees and many researchers agree on this. This was the reason we decided to adapt the general list of TIS building blocks (see section 2) by including 'safety' as a separate building block. In other industries 'safety' may be seen as a kind of cost-benefit factor but in the maritime industry safety is a separate factor. What the interviewees cannot agree on is whether MASS will increase safety onboard or not [21-24]. The idea that MASS has the potential to reduce accidents onboard is widespread, still; few see this potential currently realized. Therefore, the building block is extremely important, but only partially present [10].

Summarizing the above analysis, we show the status of the TIS building blocks for MASS in Table 2 below.

Table 2: MASS Building blocks

MASS building blocks		
Quality & Performance	Very Important	Partially Present
Cost-Benefit Ratio	Extremely Important	Absent
Production System	Moderately Important	Present
Network formation & Coordination	Slightly Important	Partially Present
Customers	Extremely Important	Partially Present
Innovation Specific Institutions	Extremely Important	Partially Present
Safety	Extremely Important	Partially Present

After the analysis of the building blocks, we further investigated the factors that influence the presence of these building blocks aiming at understanding the causes of the factors that currently influence the adoption of MASS.

4.1.8. *Knowledge and awareness of technology*

As a technology that is not yet fully proven, knowledge about MASS technological principles is not yet fully mastered. Some operational aspects of MASS, such as the application of Collision Regulation (COLREG) rules issued by the IMO, are not yet grasped by AI systems, given it was made to be used by sailors. This opens room for interpretation. Communication difficulty between vessels and harbors or between conventional vessels and autonomous vessels was also pointed out by many interviewees as a difficult operational aspect to be solved. For lower levels of autonomy, where the crew is still onboard or when an operator is in the control room these aspects do not pose a problem, as the operator can correct possible AI mistakes. But it is an important influencing factor for unmanned autonomy. The lack of deep understanding of MASS technological principles is also a barrier to the regulators, who do not know enough about the advantages of disadvantages of the technology to regulate its use. This influencing factor is, then, partially hampering MASS adoption in the current MASS TIS [10].

4.1.9. *Knowledge and awareness of application and market*

Our interviewees noted that knowledge about the MASS application is a critical aspect, highlighting that even in the literature there are still misunderstandings about the difference between an autonomous and unmanned vessel, for example. Because of this reoccurring misunderstanding, many seafarers are afraid of being replaced by a machine. Ship owners are also not aware of how MASS can contribute to their operations. Most of the news about autonomous vessels is about unmanned autonomous projects, reinforcing the view that autonomous vessels are unmanned. Technology developers and researchers also lack knowledge about the maritime industry, disregarding the peculiarities of the industry such as its conservative thinking and the seafarers' knowledge as technology operators. Despite an increase in knowledge about MASS application in recent years, this influencing factor is still considered a partial barrier [10].

4.1.10. *Natural, human, and financial resources*

From the perspective of the ship owners, visualizing a good cost-benefit ratio is still difficult without the guarantee of crew reduction, and without this clear financial benefit, most companies do not want to have MASS only to increase safety. In this context, the problem seems to be the allocation of resources, rather than its availability.

The education of seafarers was also mentioned as a necessary aspect of MASS operation. The maritime schools follow IMO guidelines regarding the curriculum and as IMO has not yet certified this technology, learning how to

work with MASS is not yet in the curriculum. This aspect does not have a large influence because professional courses can be offered to increase the knowledge of officers after their graduation, as is currently done with Dynamic Positioning Officers.

The shortage of qualified seafarers [25,26] to operate the vessels is an encouraging factor for the development and adoption of MASS. Commercial vessels can enhance their profit margin, and the navy can maintain the vessels' operations even when there is little public interest in working for the military.

None of the interviewees mentioned natural resources as an influencing factor. The shortage of human resources has a positive influence on the adoption of MASS. Both the availability of resources and the lack of qualified seafarers influence MASS adoption.

4.1.11. Competition

At the time of the data collection, not many companies we interviewed were working on MASS solutions and very few could offer a complete integration of maritime equipment and autonomy. This could be one of the reasons companies offering MASS solutions do not experience competition among themselves, but with the status quo, namely the conventional ships. MASS can make barge transport cheaper and more efficient, saving fuel costs and reducing waiting time, competing with truck transport. The current focus on CO_2 emissions challenges truck transportation and creates an ideal window to make this modal shift to barges. However, when trucks become electric, and later autonomous, barges will most likely have reduced commercial value.

Even though many international companies offer MASS solutions to the market, the competition is still too small to have a positive influence on the price and quality. Consequently, MASS is still not able to compete on equal terms with the status quo, either in the form of conventional vessels or in a modal shift, which deems the competition a partial hampering factor.

4.1.12. Macroeconomic and strategic aspects

The maritime industry was recently at the bottom of its ordering cycle, and it will go further up in the coming 10 years, so more ships are expected to be ordered and built in the near future. The decision of which technology to add to a newly built ship considers an operational asset life of 20 years, ideally with as few modifications as possible. The lack of regulations about MASS increases the financial risk of investment for ship owners, which is an important aspect of the low-profit-margin risk-averse maritime industry. The Dutch Navy has more flexibility regarding national and international regulations and is not pressured by profit margin, decreasing the risk of the investment. This is similar for navies in other countries. The ordering cycle stage and the reduced naval risks encourage both the military and the commercial vessels to invest in new technology. The difference is the level of risk each niche of the maritime sector needs to accept to design a strategy that suits the current macroeconomic scenario. Therefore, macroeconomic and strategic aspects represent an encouraging factor, only depending on the risk inclination of the investor.

4.1.13. Socio-cultural aspects

The maritime industry is not known for its openness to change; it is a conservative industry with a very strict task division on board. Society at large also influences the adoption of autonomous technologies, and people want the technology to be safer than humans; a mistake from a machine that ends up in the loss of human life is unacceptable, despite its frequency. Besides the overall fear of AI, many seafarers worry about their job and their safety, a social aspect of MASS that is also related to the knowledge and application of the technology. While society at large plays a role, ultimately the crew members are paramount to MASS acceptance; the crew operates the system, therefore if the members do not operate it correctly, the technology will not deliver its expected benefits. Benefits not delivered affect performance, which can affect adoption and stop (or delay) diffusion. Therefore, the socio-cultural influencing aspect is judged as a partially hampering factor.

4.1.14. Accidents and Events

National and international disasters and events have historically changed maritime regulations, such as the current international regulations which were developed after tragic accidents like the sinking of the Titanic, and the oil spill caused by the Exxon Valdez [27]. Accidents, however, are not unusual at sea; 65.8% of the accidents in the maritime industry are caused by human errors and MASS could decrease these numbers [28,29]. Despite not having one accident to name, the large number of accidents involving ships can be characterized as an encouraging factor to the adoption of MASS, given one of the selling points of the technology is to reduce accidents.

4.1.15. Complementary Products & Services

The Ortt & Kamp framework [11] introduces complementary products and services as a TIS building block, which is a factor capable of influencing the adoption

of a breakthrough technology on its own. Complementary products & services for MASS were identified as reliable internet, a standard for data transfer and differentiated insurance policies for vessels equipped with MASS [10]. None of these, however, is seen by the interviewees as capable of affecting MASS adoption directly. Therefore, for MASS we consider them as influencing factors instead of a TIS building block. According to our interviewees, differentiated insurance policies, in the form of a complementary service, can influence the cost-benefit ratio, a building block that is paramount for the adoption and subsequent diffusion of MASS [10]. Data Transfer Standards and 100% internet coverage can also improve technical performance, influencing the quality & performance building block. Given its absence, Complementary Products and Services are a hampering influencing factor in the MASS TIS [10].

4.1.16. Operational Aspects

Operational aspects are not mentioned as influencing factors in Ortt & Kamp's framework [11] but were mentioned by many interviewees. These aspects refer to factors that influence the regular operation of autonomous vessels, such as maintenance in the propulsion engines, especially diesel engines, cargo value and inspection, communication with other vessels and port facilities and finally stormy seas. These aspects are related to the unmanned operation of the vessels because as long as the vessel is locally operated, there will be a crew on board to give maintenance to the engines, ensure cargo safety, communicate, and perform maneuvers to ensure the vessel's water tightness. Consequently, this factor does not have an immediate influence on MASS adoption, but it will have an effect later in the large-scale diffusion of MASS if/when vessels become operated by Remote Operation Centers (ROCs).

The status of the different influencing factors for MASS is shown in Table 3 below.

Table 3: MASS Building blocks

MASS Influencing Factors	
● Encouraging Factor ● Partial Barrier ● Barrier	
Knowledge and awareness of technology	●
Knowledge and awareness of application and market	●
Natural, human, and financial resources	●
Competition	●
Macroeconomic and strategic aspects	●
Socio-cultural aspects	●
Accidents and Events	●
Complementary Products & Services	●
Operational Aspects	●
MASS Building Blocks	
● Present ● Partially present ● Absent	
Quality & Performance	●
Cost-Benefit	●
Production System	●
Network formation & Coordination	●
Customers	●
Innovation Specific Institutions	●
Safety	●

4.2. RQ2: How can potential (niche introduction) strategies be derived from the analysis of the TIS around MASS?

Now that we have knowledge of the status of the different TIS building blocks for MASS and how the influencing factors influence that status, we can derive niche strategies for the introduction of MASS in the market. The logic here is that the niche strategies are specifically derived to address incomplete or partially complete TIS building blocks in combination with their influencing factors.

4.2.1. Quality & Performance

The main issue with MASS lies in the fact that the technology is not yet proven. Little is known about commercial tests as companies only share Public Relations articles when they have positive outcomes. Most interviewees indicated that they did not know much about how MASS works.

Demonstrating, however, is not impossible and it was suggested by crew members interviewed as a way to prove MASS can deliver what it promises. The Demo experiment and Develop Niche strategy, explained in the theory

section, suggests a similar approach. In this niche strategy, the technology developers demonstrate the product in a controlled environment to further develop it and learn with it at the same time it enhances the awareness of the public [15]

Given the nature of MASS, it needs a demonstration that can be followed at a distance, like the blog created by the Mayflower developers [30] where it was possible to follow the vessel online, demonstrating its capabilities on a daily basis. Such live demonstrations can enhance trust in the system's capabilities. As a by-product, such online activity might also improve awareness of the technology and its application. To ensure that the demonstration can only take place digitally, this niche strategy is named the Digital Demo experiment and Develop Niche Strategy.

4.2.2. Cost-Benefit Ratio

One point was very clear for most interviewees: If the MASS suppliers want to sell, they need to present a clear business case and therefore a positive cost-benefit ratio. From the perspective of the companies offering technology, regulations must recognize MASS and allow crew reduction. If crew reduction is approved, the decision to adopt MASS becomes clear-cut for ship owners. IMO is working on a MASS code, with the non-mandatory part expected to be ready in 2025 and the mandatory code to enter force in 2032. Two strategies that can be used to overcome the cost-benefit ratio barrier, both described in the theory section [15]. The first one is the Top Niche strategy, which aims at a market segment that is not constrained by the price of the technology. The Royal Dutch Navy is more interested in crew reduction than in cost-benefit ratio and it does not have the obligation to comply with international rules regarding crewing [10]. Therefore, it is a suitable option for the Top Niche strategy. According to our interviewees, another maritime market segment known for its lack of budget constraints is the leisure yacht market segment, which can be open to MASS adoption without cost-benefit concern.

The second one is the Hybridization Niche Strategy, which entails making the new product compatible with existing complementary services or combining the new product with the old version [15]. A similar strategy is being pursued by two companies interviewed. Both offer MASS to existing vessels, while one ensures it sells a product and not a project, offering installation in every ship. The other focuses on retrofitting or installing the equipment in existing and operating vessels. Currently, the second company offers operational visibility to ship owners, not autonomy. The equipment installed, however, will learn with the vessel and, in the future, the company intends to add autonomy to the already sold product. The approach used by the second company resembles a combination of the previously mentioned Hybridization Niche Strategy with the Lead User Niche Strategy, which is explained is as a partnership with users that are willing to co-develop and experiment with technology [15].

4.2.3. Network Formation & Coordination

According to our interviewees, there is little collaboration between MASS providers and the network of complementary services, which directly affects the knowledge of the technology and its application. Interestingly, our interviewees understand the network as the exchange of information and ideas between the industry and its players. In the MASS context, a university of technology studying MASS technology could fulfil the public innovator role in the system. Despite not having control over what the researchers develop in the universities, companies offering MASS solutions could profit from a close relationship with academic researchers and enhance network formation for MASS. In general, research institutes can play an important coordinating and even leading role in the initial stages of development and diffusion of a radically new technology (e.g., [31])

4.2.4. Customers

Without customers, there is no diffusion and to purchase, customers need to be aware of the technology and have the resources to purchase it [11]. We found that potential customers lack awareness of MASS competencies, which affects their interest in purchasing and installing the technology in their vessels. To increase customers' interest, companies offering MASS solutions need to educate their market.

The Educate Niche Strategy is recommended to ensure customers have a better understanding of the technology [15]. The authors suggest organizing conferences and closing partnerships with universities to enhance the knowledge and awareness of innovative technologies. Organizing conferences can be rather costly to companies struggling with financial resources, and this is the situation for most small companies offering MASS solutions.

Another option to increase technology awareness is to partner with universities in projects related to MASS or make active use of the already existing autonomous vessel programs funded by national governments and institutions. Participation in these projects can have a twofold contribution to MASS adoption. On the one hand, smaller companies can piggyback on the resources of national programs to enhance awareness about their product. On the other hand, the participation of informational intermediaries, the universities and the joint industry projects, facilitates the network formation for MASS. Therefore, adopting the Educate Niche Strategy can help the companies solve the Customers' barrier to MASS adoption and diffusion.

4.2.5. Innovation Specific Institutions

Institutions, more specifically regulations regarding the number of crew members onboard, pose a large barrier to MASS adoption. The uncertainty about crew reduction possibilities hampers the adoption of MASS, except for military applications. Therefore, changes in regulations can have a substantial influence on MASS diffusion.

The Geographic Niche strategy is proposed when breakthrough technologies are restricted by regulations, suggesting introducing the technology in areas with more flexible rules [15]. Considering the power of the flag state to determine the number of crew members in vessels flying their flag and the concept of territorial waters [32], companies offering MASS can request a Minimum Safe Manning reduction with the flag state, and within territorial waters. Despite not having control of the results of these crew reduction requests, the companies offering MASS solutions stand better chances of success focusing on national entities, a solution that was also proposed by Rødseth [23] than waiting for an IMO international recognition.

4.2.6. Safety

Safety is not a building block encompassed by the Ortt & Kamp framework [11], although it is paramount to MASS adoption [10]. Therefore, there is no entry strategy presented in our theory section designed to ensure the safety of a product in its market introduction.

Wróbel, Montewka, & Kujala [25] highlight the complexity of proving safety. Safety can be proven in different ways, for example by meeting requirements, by testing designs or by measuring actual accidents after implementation. All these ways have their shortcomings and are incomplete in terms of ensuring safety. According to Ventikos, Chmurski, & Louzis [33] the autonomous shipping systems should mimic human logic and actions in their decisions. For example, when an autonomous ship gives way to another ship, the maneuver should be clear and visible, rather than only optimal. Although helpful, the authors' suggestion does not influence MASS adoption because the input of humans in the AI decision-making process is not visible to crew members and ship owners that will operate and buy the technology respectively. Therefore, this study cannot propose a useful strategy yet to address safety concerns. The table with the main strategies that we found for the main barriers is shown in Table 4 below.

Table 3: Strategies to Surmount MASS Barriers

Strategies to Surmount MASS Barriers		
● Present	● Partially Present	● Absent
Quality & Performance	●	Digital Demo experiment and Develop Niche Strategy.
Cost-Benefit Ratio	●	Top Niche Strategy.
		Hybridization Niche Strategy
		Lead User Niche Strategy
Production System	●	
Network formation & Coordination	●	Build relationships with institutes researching MASS
Customers	●	Educate Niche Strategy
Innovation Specific Institutions	●	Geographic Niche Strategy
Safety	●	… (No strategy defined yet)

5. Discussion

In this article, we focus on MASS and in order to illustrate how to analyze the wider system around a radically new technological innovation. This analysis allows us to explore an upcoming market and derive introduction strategies into that market.

5.1. Answers to the research questions

Our first research question is 'What is the status of the building blocks that together form a Technological Innovation System (TIS) around MASS and how is that status affected by influencing factors?'. Our analysis revealed that the main barrier for MASS is the lack of cost-benefits, or a lacking viable business model to apply MASS. Other building blocks are partly complete except for the production system for MASS technology, which was found to be ready and complete. So, only one building block of the TIS is considered complete and all the other ones are incomplete. That means that the TIS analysis reveals that MASS is not even close to the start of large-scale diffusion. Similarly, several influencing factors were found to have a negative effect, among which are operational aspects and a lack of complementary products and services. The status of the influencing factors confirms the idea that MASS will not start to diffuse on a large scale soon, as also shown in [34].

Our second research question is 'How can potential (niche introduction) strategies be derived from the analysis of the TIS around MASS?'. The strategies presented in this article show different ways to deal with each incomplete or absent building block in the MASS TIS. Depending on the influencing factor that is found to be the cause of the incomplete building block, different niche strategies can be adopted. The lack of quality and performance of the technology and the lack of knowledge of how to apply the technology in practice, for example, can both be improved with a Demo Experiment and Develop Niche strategy. In this niche strategy, the technology developers demonstrate the product in a controlled environment to further develop it and learn to use it and at the same time it enhances the awareness of the public [15]. The biggest barrier for MASS turned out to

be the lack of a good business model, referred to as the cost-benefit ratio barrier. Two strategies are suggested to overcome the cost-benefit ratio barrier. The first one is the Top Niche strategy, which aims at a market segment that is not constrained by the price of the technology. An example of a niche segment in which this strategy may work is the Royal Dutch Navy, which is more interested in crew reduction than in financial costs. Another segment in which this strategy may apply is the high-end leisure yacht market segment. The second strategy to overcome the cost-benefit barrier is the Hybridization Niche Strategy, which entails making the new product compatible with existing complementary services or combining the new product with the old version. In a similar way, several other specific niche introduction strategies are suggested to deal with incomplete building blocks, or barriers in the market.

5.2. Contributions

The article contributes to the scientific domain of TIS in several ways. Firstly, the article presents an adapted TIS framework, specifically reflecting the industrial context of international cargo shipping. This focuses attention on the notion that a general framework needs to be modified for the context in which it is used. We also found several such modifications necessary in studies adapting the same general TIS to other industrial contexts (e.g., [35-36]). One of the adaptations made to the TIS framework in this article is that in the MASS, safety is considered a building block instead of an influencing factor. We found it to be affected by the influencing factors accidents and events, knowledge and awareness of the technology and socio-cultural aspects. With regard to strategy for this building block, it is possible to infer that a strategy to overcome the safety barrier must include measures related to the influencing factors accidents and events, socio-cultural aspects (the norms and values of behaviors at sea) and awareness of the technology. However, this study has not found any suitable strategies yet to overcome this barrier.

Secondly, another way to reflect the intricacies of a particular context is to indicate how important each TIS building block's status is. In this article, this has been done for the first time in a TIS-analysis. Building blocks represent aspects that generally, across a wide range of industrial and geographic contexts, are important for large-scale diffusion of a technological innovation. However, depending on the geographic or industrial context, their importance may vary considerably. In the case of MASS we found that the only complete building block was considered just moderately important, for example. In contrast, the building block that was fully incomplete, however, was considered extremely important by interviewees. This extra analysis in more detail confirms the overall conclusion that MASS is far from large-scale diffusion.

The article also represents managerial and societal contributions. The framework is a kind of dashboard that, when tracked carefully over time, indicates when MASS may begin to diffuse on a large scale. That is an important managerial instrument because it allows companies to prepare for upcoming large-scale diffusion. Conversely, the framework can also prevent unnecessary preparation for introducing or adopting MASS when the conditions for large-scale diffusion are simply not met by far. Societally, the analysis using the TIS framework can help governments and companies to address barriers, remove them and hence speed up the development and start of large-scale diffusion.

5.3. Limitations

The first limitation is that the research is time-constrained and in view of the speed of development of improved technological systems, the results can be outdated within a few years. On the other hand, some radically new technologies seem to remain in a status that is considered to be “just before large-scale diffusion” for decades. Inevitably the analysis needs to be updated in the future to explore whether the status of building blocks and influencing factors have changed. As a result of such changes, new niche introduction strategies may need to be considered.

A second limitation is that interviewees came from one country, the Netherlands. The reason is that we had access to the experts, researchers and practitioners in this country. The effect of this limitation may be limited because all interviewees showed themselves to have an international perspective.

The TIS framework considers the building blocks as separate, independent aspects. That is a third limitation. Actors and factors in a TIS are interrelated, which means that adopting a strategy to deal with a barrier might create a new barrier and hence the strategy may not have the desired effect. An example can be taken from the cost-benefit ratio building block and the Top Niche Strategy. If a company offering MASS solutions focuses only on this strategy as the path to diffusion, it has little chance of succeeding, as both the navy and the leisure yachts market segments are only a small percentage of the maritime sector. The strategy would be minimizing the importance of the customers, safety and institutions, all deemed to be also extremely important in the MASS TIS. Since many factors are considered extremely important, a pathway addressing them simultaneously has a higher probability of success on the road to MASS diffusion.

5.4. Future research

In future research a wider set of interviewees, from across different countries, can be approached. This may reveal different conditions in different regions, and how

these conditions call for different niche introduction strategies.

We used qualitative data and analysis methods. In future work it may be useful to track the changes of the building blocks and influencing factors over time. That may call for a quantification of measuring these building blocks and influencing factors. Using the current qualitative approach, we could only explain why a building block changes from red (a complete barrier) to orange or green (no barrier). Such a crude and qualitative assessment may suffice to reflect a current condition and choose a niche introduction strategy but will not suffice to carefully track changes over time.

Then there is the issue of safety concerns. As this building block was newly added to the TIS framework in this study, there is no suitable niche strategy yet to address this. This is another interesting topic for future research.

The interrelation of actors and factors in a TIS calls for more advanced models taking all the building blocks and influencing factors into account. The current TIS analysis represents an overview that reflects conditions at one point in time and ignores possible interactions. This suffices to select possible niche introduction strategies that fit these conditions. Inevitably the analysis using the TIS framework needs to be repeated over time to reflect changes in the conditions. If interrelationships are taken into account, then more dynamic models could be developed that also capture the causes of changes over time. Studying the development and diffusion of radically new technological innovations, it becomes clear how fascinating the dynamics in these processes can be.

6. Conclusion

Large-scale application of Maritime Autonomous Surface Ships (MASS), being a radically new technological innovation, is still hampered by several barriers. In this paper we have applied the TIS framework to identify the main barriers and their causes and to formulate niche strategies. The main barriers include the cost-benefit ratio, safety issues and quality and performance. Niche strategies such as the Top Niche Strategy can be used to introduce MASS in niche markets first, for example the navy. This could be a first step towards the introduction of this technology on a large scale.

References

[1] M. L. Tushman, P. Anderson, "Technological discontinuities and organizational environments," *Administrative science quarterly*, vol. 31, no.3, pp. 439-465, 1986.

[2] R. Garcia, R. Calantone, "A critical look at technological innovation typology and innovativeness terminology: a literature review," *Journal of Product Innovation Management*, vol. 19, no. 2, pp. 110-132, 2002.

[3] R. D. Dewar, J. E. Dutton, "The adoption of radical and incremental innovations: an empirical analysis," *Management Science*, vol. 32, no. 11, pp. 1422-1433, 1986.

[4] S. P. Schnaars, Megamistakes. Collier Macmillan, 1989.

[5] R. Agarwal, B. L. Bayus, "The market evolution and sales takeoff of product innovations," *Management Science*, vol. 48, no. 8, pp. 1024-1041, 2002.

[6] J. R. Ortt, "Understanding the Pre-diffusion Phases," in Gaining Momentum, Imperial College Press, pp. 47–80, 2010.

[7] B. Carlsson, R. Stankiewicz, "On the nature, function and composition of technological systems," *Journal of evolutionary economics*, vol. 1, pp. 93-118, 1991.

[8] A. Bergek, "Technological innovation systems: a review of recent findings and suggestions for future research," *Handbook of sustainable innovation, vol*. 30, pp. 200-218, 2019.

[9] M. Kurtinaitis Joukes, J. R. Ortt, M. de Bruijne, "Autonomous shipping: current status and main barriers to large-scale diffusion," (Unpublished, 2023).

[10] M. Kurtinaitis Joukes, J. R. Ortt, M. de Bruijne, "A system's perspective analysis of barriers to MASS large-scale diffusion", *Proceedings of the 29th International Conference on Engineering, Technology, and Innovation: Shaping the Future, ICE 2023. IEEE*, 2023.

[11] J. R. Ortt, L. M. Kamp, "A technological innovation system framework to formulate niche introduction strategies for companies prior to large-scale diffusion," *Technological Forecasting and Social Change*, vol. 180, no. 121671, pp. 1-17, 2022.

[12] E. M. Rogers, Diffusion of Innovations, The Free Press, 2005.

[13] J. R. Ortt, J. P. L. Schoormans, "The pattern of development and diffusion of breakthrough communication technologies," *European Journal of Innovation Management*, vol. 7, no. 4, pp. 292–302, 2004.

[14] H. Markard, M. Hekkert, S. Jacobsson, "The technological innovation systems framework: Response to six criticisms," *Environmental innovation and societal transitions*, vol. 16, pp. 76-86, 2015.

[15] J. R. Ortt, D. J. Langley, N. Pals, "Ten niche strategies to commercialize new high-tech products," *2013 International Conference on Engineering, Technology and Innovation (ICE) & IEEE International Technology Management Conference*, pp. 1-12, 2013.

[16] M. Kurtinaitis Joukes, "The Diffusion of Maritime Autonomous Surface Ships," (TU Delft repository, 2012). http://resolver.tudelft.nl/uuid:b0bfb05f-83d2-4a47-a28d-bdb067793a3d, 2021.

[17] M. N. K. Saunders, P. Lewis, A. Thornhill, Research methods for business students, Pearson Education, 2015.

[18] E. M. Rogers, "The diffusion of home computers among households in silicon valley," *Marriage Fam. Rev.*, vol. 8, no. 1–2, pp. 89–101, 1985.

[19] G. Papachristos, G. van de Kaa, "A system dynamics model of standards competition," *IEEE Transactions on Engineering Management*, vol. 68, no. 1, pp. 18-32, 2020.

[20] M. A. Schilling, Strategic Management of Technological Innovation, Mc Graw Hill Education, 2017.

[21] A. Komianos, "The autonomous shipping era. operational, regulatory, and quality challenges," *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 12, no. 2, 2018.

[22] "Future proofed?," Nautilusfederation.org. [Online]. Available: https://www.nautilusfederation.org/globalassets/federation/resources/ downloads/report_auto_8.2.18.compressed.pdf. [Accessed: 27-Feb- 2023].

[23] Ø. J. Rødseth, H. Nordahl, "Definitions for autonomous merchant ships," (Unpublished, 2017).

[24] "The future of Smart Autonomy is here," Wartsila.com. [Online]. Available: https://www.wartsila.com/docs/default-source/whitepapers/smart_autonomy_e-report_2021-final.pdf?sfvrsn=31024843_3. [Accessed: 27-Feb-2023].

[25] K. Wróbel, J. Montewka, P. Kujala, "Towards the assessment of potential impact of unmanned vessels on maritime transportation safety," *Reliab. Eng. Syst. Saf.*, vol. 165, pp. 155–169, 2017.

[26] H. Leggate, "The future shortage of seafarers: will it become a reality?," *Maritime Policy & Management*, vol. 31, no. 1, pp. 3-13, 2004.

[27] Galić S, Lušić Z, Pušić D. "Seafarers market," *International Journal of New Trends in Arts, Sports & Science Education*, vol. 1, no. 3, 2012.

[28] A. A. Hebbar, "Risk communication and maritime safety legislation," (World Maritime University Dissertations, 2006), Retrieved from http://commons.wmu.se/all_dissertations/346?utm_source=commons.wmu.se%2Fall_dissertations%2F346&utm_medium=PDF&utm_campaign=PDFCoverPages

[29] "The future of Smart Autonomy is here," Wartsila.com. [Online]. Available: https://www.wartsila.com/docs/default-source/whitepapers/smart_autonomy_e-report_2021-final.pdf?sfvrsn=31024843_3. [Accessed: 27-Feb-2023].

[30] Mayflower 400. 2020. The Mayflower Story. Retrieved from 1620-2020 Mayflower 400: Available at: https://www.mayflower400uk.org/education/the-mayflower-story/

[31] O. Dedehayir, S. J. Mäkinen, J. R. Ortt,. "Roles during innovation ecosystem genesis: A literature review," *Technological Forecasting and Social Change*, vol. 136, pp. 18-29, 2018.

[32] J. N. Mansell, Flag state responsibility: Historical development and contemporary issues, Springer Science & Business Media, 2009.

[33] N. P. Ventikos, A. Chmurski, K. Louzis, "A systems-based application for autonomous vessels safety: Hazard identification as a function of increasing autonomy levels," *Saf. Sci.*, vol. 131, no. 104919, p. 104919, 2020.

[34] C. Kooij, "Towards unmanned ships, A task analysis to identify economically viable low manned ship concepts," (Ph. D Thesis, Delft University of Technology, 2021).

[35] C. Verdegaal, "Developing Effective Strategies for the Deployment of Sustainable Aircraft Technologies by the Aviation Industry," (Thesis report MSc MOT, Delft University of Technology, 2023).

[36] M. P. Dwisatyawati, "Analysis on Mass Adoption of Solar Electric Vehicles in Indonesia. A Technological Innovation System Study," (Thesis report MSc COSEM, Delft University of Technology, 2022).

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/bysa/4.0/).

7. Appendices

7.1. Appendix 1: Descriptions of the TIS building blocks

Building Blocks	Description
Product Performance and Quality	The newly developed technological product has sufficiently good quality and performance or is expected to have it shortly when compared to competing products. If the new technology suffers from low quality and is unable to meet the customer's needs and requirements large-scale diffusion will be hampered.
Product Price	The price of a product involves not only its monetary costs but also non-financial costs such as time, effort to implement the new product, switching costs, and costs to find new suppliers. For large-scale diffusion, the price should be reasonable when compared to other competitive alternatives of the same technology.
Production System	A production system that can deliver large quantities of high-quality products is paramount to the large-scale diffusion of technological innovation. Not only creating a production system but fine-tuning it to profit from the learning effect costs time and money, which can delay the diffusion process.
Complementary Products and Services	Complementary products and services support production, distribution, adoption and finally the disposal, if necessary, of the innovative technological product. Together, the network of complementary products and services can induce other innovations and motivate companies to align their strategy, which will ensure large-scale diffusion. The lack of these products and services forms a barrier that blocks the diffusion.
Network Formation and Coordination	A supply chain network with suppliers of parts, distributors, complementary services, and other actors is important for the diffusion of an innovative product. The lack of alignment in this network can impede the large-scale diffusion of the technology.
Customers	The customers are extremely important in the TIS and the diffusion of the innovative technology, without consumers, there is no diffusion. Customers must be able to afford the product,

	understand enough the product's capabilities to have the correct understanding of its risks and benefits, as well as use it. The development of technological innovation without the perspective of the future customer often results in issues that hamper diffusion, such as lack of integration with the customer's routine of use, alignment to other technologies already in use. Some technologies have to be adjusted later to fit the customer's wishes.
Innovation-Specific Institutions	To form a TIS network, innovation-specific institutions must be present. These institutions refer to regulations, laws, standards, and government policies, which can either block or encourage the formation of the TIS. Factors such as stability of the political and legal systems, quality norms, and property rights produce trust in the system, which in turn, increases investments and facilitates the TIS formation.

7.2. Appendix 2: Descriptions of the influencing factors

Influencing Factors	Description
Knowledge and Awareness of Technology	This aspect refers to both fundamental and applied knowledge. They comprise knowledge about the product itself, its production system, and complementary products, as well as the knowledge about its application, which is necessary to develop, produce, maintain, and use the product. Actors in the TIS need to be aware of both types of knowledge, which can be developed through research and practical experimentation. The process of learning about technological innovation, mainly through education, is vital to increase both types of knowledge.
Knowledge and Awareness of Application and Market	The price of a product involves not only its m This influencing factor relates to the use and applications of innovative products, as well as the market structure and its relevant actors. Lack of application knowledge can affect customers as well as suppliers. Suppliers without application knowledge are not able to target the correct customer segment, and customers without knowledge of the purpose, how to use or buy an innovative product will be a barrier to product diffusion. Market analysis, experimentation, and learning by doing are ways of developing this type of knowledge.
Natural, Human and Financial Resources	All three types of resources, natural, human, and financial resources are required for the production system of the innovation itself as well as the complementary products and services. The lack of raw material, people with the necessary competencies or capital to invest in technological innovation can hamper technology diffusion. The lack of financial capital specifically blocks the development of radically new products and their future market formation.
Competition	The presence of competition influences market formation, especially when there are alternative competing technologies requiring different components, production systems and complementary products. Such a situation leads to uncertainty, which hampers market formation, an important building block for large-scale diffusion. Competition also influences the relative price and performance, shaping the investment in production and complementary products and finally influencing customer adoption.
Macroeconomic and Strategic Aspects	The macro-economic situation can largely influence the conditions for the TIS formation. While economic growth facilitates the formation of TIS with funds for incubators, for example. An economic recession can hamper the large-scale diffusion of a breakthrough innovation given the reduction of these incentives to innovation. The strategies and country policies are often affected by macroeconomic conditions, such

	as the market structure, making both macroeconomic and strategic aspects not only dependent on each other but also of large influence in the TIS building blocks.
Socio-Cultural Aspects	The socio-cultural aspects are the norms and values held by the potential customers and stakeholders of innovative technologies. These norms are not necessarily laws, policies, or regulations, but informal common norms that impact the behavior of TIS actors, such as safety concerns. These rules change with time, and they can influence technology adoption in both ways, encouraging or blocking the technology large-scale adoption.
Accidents and Events	Events such as wars, or natural disasters, together with national or international accidents generate a great impact on the formation of TIS-es. Some of these events stimulate the formation of TIS for radically new technologies because they generate new needs, which have to be fulfilled with innovative technologies.

Received: 30 October 2024, Revised: 14 December 2024, Accepted: 15 December 2024, Online: 19 January 2025
DOI: https://dx.doi.org/10.55708/js0401001

Enhancing Python Code Embeddings: Fusion of Code2vec with Large Language Models

Long H. Ngo*, **Jonathan Rivalan**
Smile France, Asnières-sur-Seine, 92600, France
*Corresponding author: Long H. Ngo, Paris, France, Email: long.ngo@smile.fr

ABSTRACT: Automated code comprehension has recently become integral to software development. Neural networks, widely employed in natural language processing tasks, can capture the semantic meanings of language by representing it in vector form. Although programming code differs from natural language, we hypothesize that neural models can learn both the syntactic and semantic attributes inherent in code. This study presents an innovative approach to improve code representation and understanding for Python, building upon a previous work (code2vec extended with ASTminer). The novel method integrates embeddings from Large Language Models (LLMs) with code2vec vectors, aiming to align both semantic and syntactic information in code representations. We explore various fusion techniques, including simple concatenation, weighted sum, or attention-based mechanism, to combine LLM embeddings with code2vec vectors. We explore how semantic information from LLMs complements the structural information from code2vec, and discuss the potential impact of this synergy on software development practices. These findings open new directions for more accurate and adaptable code understanding models, with implications for improving documentation, code search, and overall software development efficiency.

KEYWORDS: Machine learning, Neural network, Large Language Model, Distributed representations, Code search

1. Introduction

In recent years, the field of automated code understanding has become increasingly crucial in software development. Certain problems in software development, such as assigning meaningful method names, highlight the need for concise semantic representations of code snippets. The core challenge lies in encoding code snippets in a way that captures semantically relevant information, transferable across multiple programs, and enables the prediction of properties like code labeling. This involves two key components: first, representing the code snippet in a manner that supports learning across different programs, and second, determining which parts of the representation are critical for property prediction and how to prioritize them.

This paper builds upon a previous study in [1] at the International Conference on Software Engineering and Artificial Intelligence (SEAI), by addressing its limitations and introducing novel methodology to improve code representation and search capabilities. Code representation presents challenges due to the need to capture both semantic and syntactic information in a format conducive to machine learning. The previous study, in [1], used a path-based representation technique to address two key tasks in software development: assigning semantic labels to code snippets and performing code searches. The path-based method represents a snippet by extracting paths from its abstract syntax tree (AST), capturing both the structure and semantics of the code, as demonstrated in previous research [2]. The first task focused on automating the prediction of a semantic label for a given code snippet. This is a challenging task because it requires learning complex relationships between the content of a method and a semantic label. Specifically, it involves condensing numerous expressions and statements within the method into a single descriptive label, which demands sophisticated techniques for code representation and classification [3]. The second task aimed to develop an efficient and effective search mechanism for locating code snippets based on query requirements. This is a critical need for developers, who often need to find and reuse existing code. Successful code search must be able to match a query against a large codebase and retrieve relevant code snippets. Addressing both of these tasks effectively would significantly enhance software development productivity and efficiency.

However, the work in [1] also revealed limitations, particularly in aligning the docstrings with extracted paths from the input code snippets. It might miss out on high-level semantic relationships. This challenge highlighted the need for more sophisticated approaches to code representation.

In this extended version, we present several significant advancements:

- Integration of large language models (LLMs) with code2vec to create more comprehensive code embeddings.
- Development of fusion techniques to combine semantic embeddings from LLMs with syntactic vectors from code2vec.

- Extensive experimental settings with various fusion strategies, including concatenation, weighted sum, and attention-based mechanisms.

By addressing these areas, we aim to push the boundaries of code representation, enabling more accurate and efficient automated understanding of Python code. This work not only builds upon the previous findings but also opens new directions for research in the intersection of natural language processing and programming language analysis.

The remainder of this paper is structured as follows: Section 2 reviews related work in code representation and analysis. Section 3 describes the work where code2vec was extended with ASTminer. Section 4 describes the present work, including the adaptation of code2vec for Python and the proposed fusion techniques. Section 5 discusses the implications of our findings. Section 6 concludes the paper with a summary of contributions and directions for future work.

2. Related works

The field of code representation and analysis has seen significant advancements in recent years, driven by the application of deep learning techniques traditionally used in natural language processing (NLP). While the conventional approach in NLP involves treating text as a linear sequence of tokens processed through neural networks [4, 5], this method has also been widely adopted for source code representation [6, 7, 8, 9, 10]. However, recent research has highlighted the potential benefits of leveraging the inherent structure of programming languages. In [11, 12, 13], the authors have demonstrated that structured representations can significantly enhance performance in various code analysis tasks.

The ability to predict program properties through learning from large-scale codebases has been a focus of numerous studies [6, 8, 9, 12, 14]. This capability opens up a wide range of applications in software engineering, including predicting names for program entities [11, 13, 15], code completion [16, 17], code summarization [6], code generation [18, 19, 20]. These applications, among others [21, 22], showcase the potential impact of advanced code representation techniques on various aspects of software development and maintenance, with minimal or no semantic analysis.

A significant breakthrough in this field came with the introduction of code2vec [3]. This innovative approach uses neural networks to learn distributed representations of code, addressing one of the most challenging problems in software engineering. Unlike traditional methods that rely primarily on static code features, code2vec uses a large corpus of code snippets to learn representations that capture the temporal dynamics of code changes. This results in more accurate and robust code representations.

The code2vec model consists of two main components: 1) An input encoder that converts code snippets into token sequences. These sequences are then used to train a neural network to predict the next tokens in the sequence, which is similar to a large language model. 2) A code vectorizer that generates continuous-valued vector representations from the output of the encoder.

These components work together to capture both syntactic and semantic information from the code, utilizing attention mechanisms to prioritize the most relevant aspects of each snippet. The resulting vector representations have proven effective in various software engineering tasks, including code completion, bug detection, and program synthesis. The architecture of the code2vec network is illustrated in Figure 1.

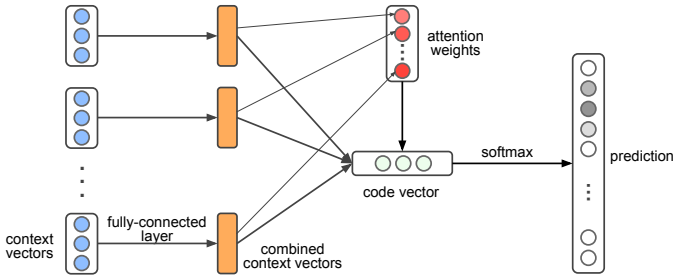


Figure 1: Original code2vec architecture. A fully-connected layer combines embeddings of each path-context. Attention weights are learned and used to compute a code vector, which predicts the label [3].

In [1], the authors adopted the code2vec neural network architecture [3], which was designed to learn low-dimensional vector representations, or embeddings, of the source code. In these vectors, the "meaning" of an entity is distributed across multiple components, allowing semantically similar entities to be mapped to nearby vectors. These embeddings enable efficient modeling of the relationship between code snippets and their semantic labels in a natural and efficient manner. Using the inherent structure of the source code, code2vec aggregated multiple syntactic paths into a single vector, leading to a more comprehensive and accurate representation of each code snippet.

The success of code2vec marks a significant departure from conventional code representation methods and has paved the way for further research in this area. Our work builds upon these foundations, extending the application of code2vec to Python and exploring novel ways to enhance its performance through integration with large language models.

3. Extended Code2vec with ASTminer for Python Code Embeddings

High-level view. In programming, a code snippet plays a pivotal role as it allows developers to write concise, reusable pieces of code applicable in different contexts. To better understand the construction of code snippets, it is essential to explore the concept of a bag of contexts. Each context within a code snippet is represented as a vector, referred to as the context vector. This vector is generated through a learning process that captures two important features of the context: (i) its semantic meaning and (ii) the attention it should receive. By aggregating these context vectors, a code vector is created, which can then be used for various downstream tasks.

3.1. Semantic labeling of code snippets

To extend code2vec for Python programming language, the ASTminer extractor [2] was employed instead of the JavaExtractor in the original code2vec [3]. ASTminer is an open source library that enables the extraction of path-based representations of codes, as shown in Figure 2. ASTminer parses the syntax tree of a code snippet to extract paths that summarize the structural and semantic essence of the code. Hence, it offers a reusable toolkit designed to simplify the task of modeling source code for machine learning algorithms, reducing the associated complexity and effort.

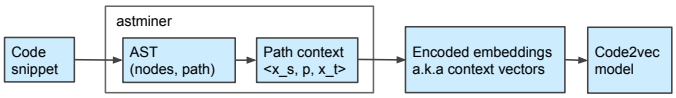


Figure 2: code2vec architecture for Python programming language.

The AST is essential in representing the syntactic structure of a program, selectively ignoring details such as punctuation, formatting or specific syntactic forms, while maintaining each node as a distinct syntactic unit. In the AST, each node represents a specific syntactic construct, such as variables, operations, or logical operators. The associated child nodes correspond to the subordinate elements related to the parent node. [2]

The ASTminer extractor uses this path-based representation to capture coding styles and structures, thereby encapsulating the logic of the code. The conversion of code into embeddings occurs in two steps: first, transforming the code into a vector, followed by combining these vectors with corresponding attention vectors. This prepares the model for further training. The architecture, depicted in Figure 3, mirrors the original code2vec embedding model.

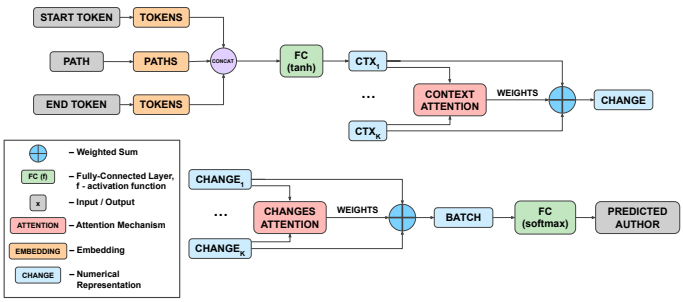


Figure 3: Overview of authorship attribution pipeline, which generates authorship-based embeddings of method changes. Furthermore, it highlights the significance of individual method changes in the authorship attribution. The method nodes and the corresponding attention weights are subsequently utilized to create author representation [23].

Following the approach in [23], the embeddings are constructed using the code and path tokens, then concatenated, and converted to a numerical representation using a fully connected (FC) activation function, as shown in Figure 3. These vectors are merged into batches where the weights of individual paths are combined, revealing the importance of each path. This pipeline is employed for predicting code titles, following the same principles as authorship attribution.

Analogous to [3], cross-entropy loss, which calculates the difference of the predicted distribution q and the "true" distribution p, is used during training phase. p assigns a value of 1 to the actual label in the training sample and 0 otherwise. Hence, the cross-entropy loss for a single sample is equivalent to the negative log-likelihood of the true label, which can be expressed as follows:

$$\mathcal{H}p\|q \; - \; _{y\in Y} py \cdot logqy \;\; -logqy_{true} \tag{1}$$

where y_{true} is the actual label of the sample. The loss is the negative logarithm of qy_{true}, the probability that the model assigns to y_{true}. As qy_{true} tends to 1, the loss approaches zero. The further qy_{true} goes below 1, the greater the loss becomes. Thus, minimizing this loss is equivalent to maximizing the log-likelihood that the model assigns to the true labels y_{true}. During training, gradient descent is utilized to iteratively update the parameters by minimizing the loss function. The learned parameters are refined using backpropagation. In other words, the parameters are iteratively adjusted by taking small steps in the direction that reduces the loss, thereby optimizing the model.

Based on the study conducted in [3], the network model can be applied effectively to predict categories for unseen code. This method involves generating a code vector by leveraging the weights and parameters learned during the training phase. The prediction is then made by determining the closest target label. This predictive model offers significant potential for improving the accuracy and efficiency of categorizing software code snippets, supporting tasks such as software maintenance, bug detection, and code optimization.

3.2. Code search

This study also builds upon previous works [24, 25, 26] to design a neural search system that uses joint embeddings of code snippets and queries. Each type of input (either code or natural language) is processed using separate encoders, trained to map inputs into a shared vector space. The system is trained to ensure that related code and queries are embedded closely in this space, enabling efficient search and retrieval of relevant code snippets. Although there are more sophisticated models that account for multiple interactions between queries and code, this architecture utilizes a single vector per query/snippet, simplifying indexing and search tasks [26]. Figure 4 provides an overview of the model architecture used for the Python language in the code search task.

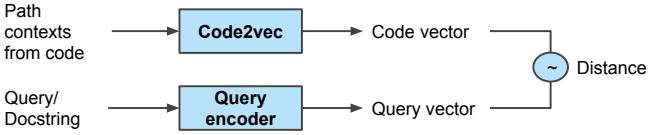


Figure 4: Text to code architecture based on code2vec model.

For this task, the extended code2vec model is used to encode Python code snippets. Queries in the form of docstrings are tokenized and passed through a query encoder, such as the Neural Bag of Words (NBoW) [27], Bidirectional RNN model [28], 1D Convolutional Neural Network [29], or Self-Attention [30]. Subsequently, the resulting token

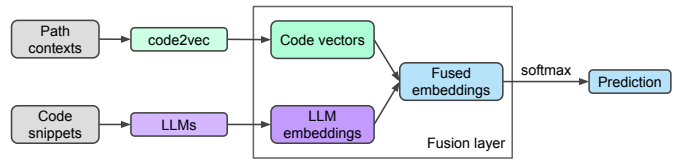
embeddings are combined into a sequence embedding using a pooling function, either mean or max pooling, and an attention-based weighted sum mechanism. The training process used a collection of N code and docstring pairs, denoted by c_i, d_i, alongside the instantiation of a code encoder E_c and a query encoder E_q. The objective is to minimize the following loss function:

$$-\frac{1}{N}\sum_i log\frac{expE_c c_i{}^T E_q d_i}{\sum_j expE_c c_j{}^T E_q d_i} \quad (2)$$

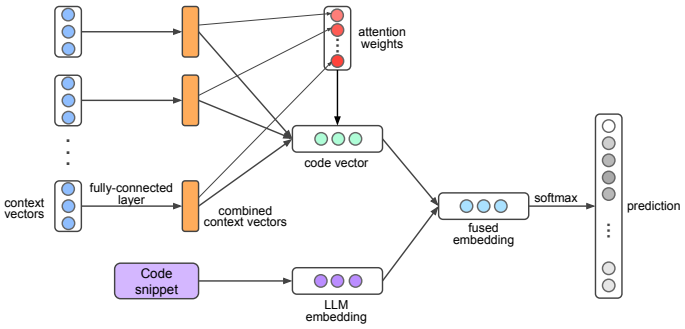
This loss encourages the inner product between matching code and docstring encodings pairs to be maximized, while minimizing the inner product between the query d_i and irrelevant code snippets c_j. During model evaluation, Mean Reciprocal Rank (MRR) is used to assess performance on the validation set. During the testing stage, the Annoy library[1], a high-speed, approximate nearest-neighbor indexing and search technique offered by Spotify, is used to index the entire CodeSearchNet Corpus. The index encompasses all functions in the corpus, including those without accompanying documentation. Annoy helps to achieve efficient and accurate retrieval [26].

4. Extended work: Code2vec and Embeddings Fusion

In this study, we propose a novel approach for enhancing the performance of code representation models by leveraging large language models (LLMs), such as Code Llama [31], Qwen2.5-coder [32], and code2vec for Python language. The integration of LLMs with traditional code representation techniques presents a promising way to enhance code understanding. Our approach combines the semantic richness of LLM embeddings with the structural insights of syntactic code2vec vectors. This fusion aims to create a more comprehensive and nuanced representation of code snippets. Figure 5 depicts the architecture of the embeddings fusion approach adapted with the code2vec model.



(a) Overview of code2vec and feature fusion approach.



(b) Updated code2vec architecture.

Figure 5: Proposed architecture of code2vec with fused embeddings using LLMs.

[1]https://github.com/spotify/annoy

We utilize Code Llama - 7B [31], a variant of the Llama model fine-tuned on code, to generate semantic embeddings. Code Llama represents the cutting edge of publicly available large language models (LLMs) for code-related tasks. It offers the potential to enhance developer workflows by improving efficiency and speed, while also reducing the barriers for individuals new to programming. Additionally, Code Llama could serve as a valuable tool for productivity and education, assisting programmers in producing more reliable and well-documented software.

We used the CodeSearchNet corpus for training and evaluation. This dataset contains about 1.1 million Python functions, spanning diverse domains and programming paradigms. Each code snippet is processed through Code Llama to produce a high-dimensional vector capturing contextual and semantic information. Concurrently, we process the same code snippets through our adapted code2vec model for Python, which generates vectors representing the structural and syntactic features of the code.

To achieve this fusion, we explore various techniques to integrate LLM embeddings and code2vec code vectors into a single, unified representation. These techniques include:

1. Simple concatenation: A straightforward approach where Code Llama embeddings and code2vec vectors are simply combined end-to-end. This method preserves all information but increases dimensionality.

$$fused_vector \;\; llm_embedding; code2vec_vector \quad (3)$$

2. Weighted sum: We apply learnable weights to each embedding type before summation, allowing the model to adjust the importance of semantic versus syntactic information.

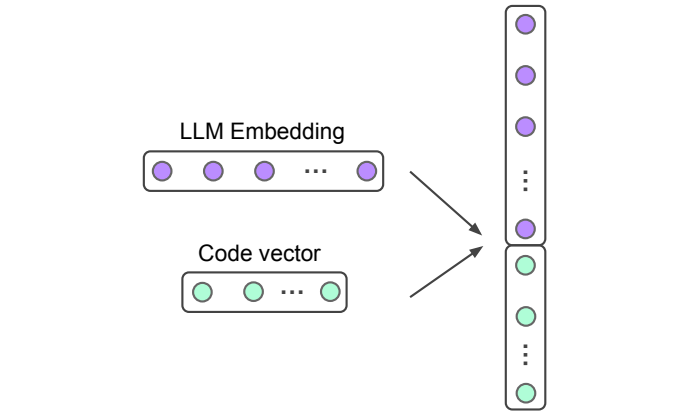
$$fused_vector \;\; w1 * llm_embedding \;\; w2 * code2vec_vector \quad (4)$$

3. Attention-based mechanisms: Inspired by transformer architectures, we implement a multi-head attention mechanism. This allows dynamic focus on different aspects of each embedding based on the specific code context. On the other hand, it allows the model to dynamically weight semantic and syntactic information based on task relevance, resolving the issue of fixed representation alignment by enabling the model to prioritize critical information.

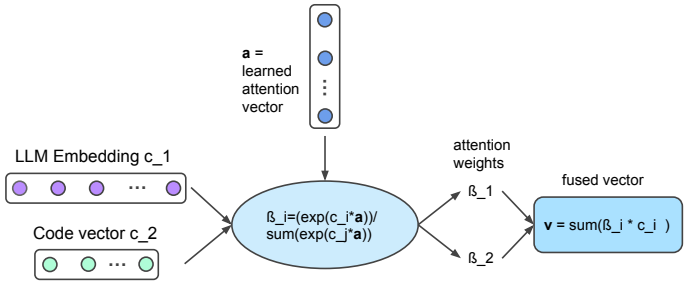
$$attention_output \;\; MultiHeadAttention llm_embedding, code2vec_vector$$
$$fused_vector \;\; FeedForward attention_output \quad (5)$$

Each method offers different advantages in terms of balancing the contribution of each type of embedding to the final representation. For instance, concatenation allows for a straightforward integration of both embeddings, while weighted sum can modulate the influence of each embedding type, and attention-based fusion offers the potential to dynamically adjust the focus on relevant features during training and inference. The fused representation is then employed in both the training and testing phases of code2vec, with the goal of improving performance on tasks such as code labeling and code search.

Figure 6 illustrates examples of two fusion techniques: concatenation and attention-based fusion. These visualizations provide insight into how each method integrates the embeddings and the potential impact on the resultant code representation.



(a) Concatenation technique.



(b) Attention technique.

Figure 6: Examples of 2 feature fusion techniques used in our feature fusion approach.

5. Discussion

The integration of LLM embeddings with code2vec vectors shows a significant theoretical advancement in code representation techniques, which offers promising directions for combining semantic and syntactic information in code understanding tasks. Here, we discuss the implications of these findings and their potential impact on the field of automated code understanding.

Our findings highlight the complementary nature of LLM embeddings and code2vec vectors, demonstrating how these two approaches can be combined to enhance code representation. LLM embeddings capture rich semantic information and contextual nuances of code, while code2vec vectors excel at representing syntactic structure. The synergy between these two types of information allows our model to develop a more comprehensive understanding of code snippets. The relevance of semantic versus syntactic information may vary depending on the specific code snippet or task at hand. The multi-head attention mechanism in our fusion approach allows the model to dynamically focus on different aspects of the code representation. This adaptability is particularly valuable for handling various programming paradigms, coding styles, or specific tasks.

The integration of LLM embeddings, which are pre-trained on vast amounts of code data, with task-specific code2vec vectors opens up new possibilities for transfer learning in code analysis. This approach allows models to benefit from the broad knowledge captured by LLMs while still maintaining the ability to fine-tune on specific tasks or codebases. This transfer learning capability could be particularly beneficial for organizations with limited labeled data or for tackling niche programming languages or domain-specific coding patterns. It suggests a path towards more generalizable code understanding models that can quickly adapt to new contexts.

The fusion of LLM embeddings and code2vec vectors also increases the computational complexity of the model. The attention-based fusion, in particular, adds a non-trivial amount of computation to the process. As we scale this approach to larger codebases or real-time applications, careful consideration must be given to balancing performance gains with computational efficiency. Future work could explore techniques for model compression or distillation to make this approach more feasible for resource-constrained environments or large-scale deployments.

The enhanced capabilities in code labeling and search have the potential to significantly impact software development practices. More accurate code labeling could lead to improved auto-documentation tools, helping maintain up-to-date and accurate code documentation. Enhanced code search capabilities could boost developer productivity by making it easier to find and reuse existing code snippets, potentially reducing duplication and improving code quality. Furthermore, these advancements could contribute to the development of more sophisticated code recommendation systems, assisting developers in writing more efficient, readable, and maintainable code.

6. Conclusion

This study has presented a significant extension to the previous work on code representation using code2vec and ASTminer for Python. By integrating large language model (LLM) embeddings with code2vec vectors, we have demonstrated a novel approach to capturing both semantic and syntactic information in code representations. While the current work focuses on architectural design and theoretical analysis, it lays important groundwork for future research in automated code understanding.

Our key theoretical contributions include: the development of fusion techniques, particularly an attention-based mechanism, to combine LLM embeddings with code2vec vectors; insights into the synergistic relationship between semantic information from LLMs and structural information from code2vec, and how this synergy can be leveraged in a novel architecture for better code understanding; theoretical foundation for improved code understanding models in future empirical studies.

These advancements address some of the limitations identified in the previous code2vec work, particularly in aligning semantic information (such as docstrings) with structural code representations. The fusion approach we developed offers a more robust and flexible way to represent code, adapting to the specific needs of different tasks and code structures.

Despite the promising aspects, several limitations and areas for further research remain:

- Language specificity: Our current work focuses on

Python. Further research is needed to assess the generalizability of this approach to other programming languages.

- Lack of empirical validation: The current study focuses on theoretical foundations and architectural design, without experimental results to validate performance claims. Comparative analysis with existing code representation techniques remains to be conducted.
- Model interpretability: While the attention mechanism provides some insight into the model's decision-making process, further work is needed to improve the interpretability of the fused representations.
- Fine-grained code understanding: Future research could explore how this approach performs on more fine-grained tasks, such as variable naming or type inference.
- Temporal aspects of code: Investigating how to incorporate information about code evolution and version history into our fused representations could provide additional valuable insights.

In conclusion, our work on fusing LLM embeddings with code2vec vectors represents a significant step forward in the field of automated code understanding. By leveraging both semantic and syntactic information, this approach opens up new possibilities for more accurate and versatile code understanding models, with potential far-reaching implications for software development practices and tools. As we continue to refine these techniques and explore their applications, we potentially anticipate a profound impact on the landscape of software development. Through rigorous experimental validation and continued refinement, these theoretical contributions can evolve into practical tools for improving code comprehension and development efficiency.

Conflict of Interest The authors declare no conflict of interest.

References

[1] L. H. Ngo, V. Sekar, E. Leclercq, J. Rivalan, "Exploring code2vec and astminer for python code embeddings", "2023 IEEE 3rd International Conference on Software Engineering and Artificial Intelligence (SEAI)", pp. 53–57, IEEE, 2023, doi:10.1109/SEAI59139.2023.10217505.

[2] V. Kovalenko, E. Bogomolov, T. Bryksin, A. Bacchelli, "Pathminer: a library for mining of path-based representations of code", "Proceedings of the 16th International Conference on Mining Software Repositories", pp. 13–17, IEEE Press, 2019, doi:10.1109/MSR.2019.00015.

[3] U. Alon, M. Zilberstein, O. Levy, E. Yahav, "code2vec: Learning distributed representations of code", *Proceedings of the ACM on Programming Languages*, vol. 3, no. POPL, pp. 1–29, 2019, doi:10.1145/3290353.

[4] S. Hu, Y. Zuo, L. Wang, P. Liu, "A review about building hidden layer methods of deep learning", *Journal of Advances in Information Technology*, vol. 7, no. 1, 2016, doi:10.12720/jait.7.1.58-63.

[5] Y. Sakai, Y. Eto, Y. Teranishi, "Structured pruning for deep neural networks with adaptive pruning rate derivation based on connection sensitivity and loss function", *Journal of Advances in Information Technology*, 2022, doi:10.12720/jait.13.1.1-7.

[6] M. Allamanis, H. Peng, C. Sutton, "A convolutional attention network for extreme summarization of source code", "International conference on machine learning", pp. 2091–2100, PMLR, 2016, doi:10.48550/arXiv.1602.03001.

[7] M. White, C. Vendome, M. Linares-Vásquez, D. Poshyvanyk, "Toward deep learning software repositories", "2015 IEEE/ACM 12th Working Conference on Mining Software Repositories", pp. 334–345, IEEE, 2015, doi:10.1109/MSR.2015.33.

[8] M. Allamanis, E. T. Barr, C. Bird, C. Sutton, "Learning natural coding conventions", "Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering", pp. 281–293, 2014, doi:10.1145/2635868.2635883.

[9] M. Allamanis, C. Sutton, "Mining source code repositories at massive scale using language modeling", "2013 10th working conference on mining software repositories (MSR)", pp. 207–216, IEEE, 2013, doi:10.1109/MSR.2013.6624029.

[10] D. Movshovitz-Attias, W. Cohen, "Natural language models for predicting programming comments", "Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)", pp. 35–40, 2013, doi:10.18653/v1/P13-2007.

[11] U. Alon, M. Zilberstein, O. Levy, E. Yahav, "A general path-based representation for predicting program properties", *ACM SIGPLAN Notices*, vol. 53, no. 4, pp. 404–419, 2018, doi:10.1145/3192366.3192412.

[12] P. Bielik, V. Raychev, M. Vechev, "Phog: probabilistic model for code", "International conference on machine learning", pp. 2933–2942, PMLR, 2016, doi:10.48550/arXiv.1602.05259.

[13] V. Raychev, M. Vechev, A. Krause, "Predicting program properties from" big code"", *ACM SIGPLAN Notices*, vol. 50, no. 1, pp. 111–124, 2015, doi:10.1145/2676726.2677009.

[14] V. Raychev, P. Bielik, M. Vechev, "Probabilistic model for code with decision trees", *ACM SIGPLAN Notices*, vol. 51, no. 10, pp. 731–747, 2016, doi:10.1145/2983990.2984041.

[15] M. Allamanis, E. T. Barr, C. Bird, C. Sutton, "Suggesting accurate method and class names", "Proceedings of the 2015 10th joint meeting on foundations of software engineering", pp. 38–49, 2015, doi:10.1145/2786805.2786849.

[16] V. Raychev, M. Vechev, E. Yahav, "Code completion with statistical language models", "Proceedings of the 35th ACM SIGPLAN conference on programming language design and implementation", pp. 419–428, 2014, doi:10.1145/2594291.2594321.

[17] A. Mishne, S. Shoham, E. Yahav, "Typestate-based semantic code search over partial programs", "Proceedings of the ACM international conference on Object oriented programming systems languages and applications", pp. 997–1016, 2012, doi:10.1145/2384616.2384698.

[18] M. Amodio, S. Chaudhuri, T. W. Reps, "Neural attribute machines for program generation", *arXiv preprint arXiv:1705.09231*, 2017, doi:10.48550/arXiv.1705.09231.

[19] Y. Lu, S. Chaudhuri, C. Jermaine, D. Melski, "Data-driven program completion", *arXiv preprint arXiv:1705.09042*, 2017, doi:10.48550/arXiv.1705.09042.

[20] C. Maddison, D. Tarlow, "Structured generative models of natural source code", "International Conference on Machine Learning", pp. 649–657, PMLR, 2014, doi:10.48550/arXiv.1401.0514.

[21] M. Allamanis, E. T. Barr, P. Devanbu, C. Sutton, "A survey of machine learning for big code and naturalness", *ACM Computing Surveys (CSUR)*, vol. 51, no. 4, pp. 1–37, 2018, doi:10.1145/3212695.

[22] M. Vechev, E. Yahav, *et al.*, "Programming with "big code"", *Foundations and Trends® in Programming Languages*, vol. 3, no. 4, pp. 231–284, 2016, doi:10.1561/2500000028.

[23] V. Kovalenko, E. Bogomolov, T. Bryksin, A. Bacchelli, "Building implicit vector representations of individual coding style", "Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops", pp. 117–124, 2020, doi:10.1145/3387940.3391461.

[24] X. Gu, H. Zhang, S. Kim, "Deep code search", "Proceedings of the 40th International Conference on Software Engineering", pp. 933–944, 2018, doi:10.1145/3180155.3180167.

[25] B. Mitra, N. Craswell, *et al.*, "An introduction to neural information retrieval", *Foundations and Trends® in Information Retrieval*, vol. 13, no. 1, pp. 1–126, 2018, doi:10.1561/1500000061.

[26] H. Husain, H.-H. Wu, T. Gazit, M. Allamanis, M. Brockschmidt, "CodeSearchNet challenge: Evaluating the state of semantic code search", *arXiv preprint arXiv:1909.09436*, 2019, doi:10.48550/arXiv.1909.09436.

[27] I. Sheikh, I. Illina, D. Fohr, G. Linares, "Learning word importance with the neural bag-of-words model", "Proceedings of the 1st Workshop on Representation Learning for NLP", pp. 222–229, 2016, doi:10.18653/v1/W16-1626.

[28] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches", "Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation", 2014, doi:10.3115/v1/W14-4012.

[29] K. Yoon, "Convolutional neural networks for sentence classification [ol]", *arXiv Preprint*, 2014, doi:10.48550/arXiv.1408.5882.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, vol. 30, 2017, doi:10.48550/arXiv.1706.03762.

[31] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez, *et al.*, "Code llama: Open foundation models for code", *arXiv preprint arXiv:2308.12950*, 2023, doi:10.48550/arXiv.2308.12950.

[32] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Dang, *et al.*, "Qwen2. 5-coder technical report", *arXiv preprint arXiv:2409.12186*, 2024, doi:10.48550/arXiv.2409.12186.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).



LONG H. NGO has done his B.E. degree in electrical and electronics engineering from Ho Chi Minh City University of Technology in 2016. He has received his master's degree in imaging and networks from Université Sorbonne Paris Nord, France, in 2017. He has completed his PhD degree in machine learning and image processing from Université Sorbonne Paris Nord, France, in 2021.

His research interests include machine learning, image processing, computer vision, and NLP.



JONATHAN RIVALAN is R&D manager at Smile, French open source services integrator based in various locations in Europe. His main line of work targets workflow automation, digital infrastructures optimization and advanced HCI developments. He is currently participating and leading national and European R&D programs, which results are available as open source solutions, such as Rating Operator (metrics transformation towards KPIs) and Palindrome.js (multidimensional monitoring for distributed systems).