JOURNAL OF ENGINEERING RESEARCH & SCIENCES

JENRS

www.jenrs.com
ISSN: 2831-4085

Volume 4 Issue 6
June 2025

EDITORIAL BOARD

Editor-in-Chief

Editorial Board Members

Prof. Hamid Mattiello
Department of Business and Economics, University of Applied Sciences (FHM), Germany

Dr. Deepak Bhaskar Acharya
Department of Computer Science, The University of Alabama in Huntsville, USA

Dr. Ali Golestani Shishvan
Department of Electrical & Computer Engineering, University of Toronto, Canada

Editorial

With growing global challenges in environmental management, healthcare, and computational sciences, innovative research continues to shape more effective and intelligent systems for detection, prediction, and decision-making. This issue showcases cutting-edge studies that leverage machine learning and numerical methods to enhance wildfire prediction, medical diagnostics, and the precision of numerical solutions. Together, these contributions reflect the critical role of computational tools in supporting urgent societal and scientific needs.

Accurate classification of wildfire types is increasingly vital in the face of rising fire incidents linked to climate change and anthropogenic pressures. A comparative evaluation of supervised machine learning algorithms applied to satellite-based environmental data identifies the Decision Tree (DT) model as the most effective classifier, with a top accuracy of 96.69% across all performance metrics. Closely following are Random Forest (RF) and Gradient Boosting Classifier (GBC), both achieving consistently high results. In contrast, Support Vector Classifier (SVC) and Logistic Regression (LR) exhibit reduced precision and F1 scores, making them less suitable for this task. By applying a robust machine learning framework to real-world U.S. wildfire datasets, the study provides actionable insights into model selection for early warning systems, ultimately supporting more responsive and informed disaster management strategies [1].

Understanding the approximation errors in numerical solutions of differential equations is critical for ensuring mathematical accuracy in engineering and scientific modeling. This study enhances the precision of error estimation by utilizing the moving nodes method, which calculates approximation errors at specific nodal points within a defined grid. By expressing the discrete solution analytically and integrating the step size hhh and accuracy order ppp, the approach provides a more accurate representation of how the numerical solution diverges from the exact one. This refinement in approximation error analysis contributes to improved reliability in simulations and numerical computations, particularly in fields where precision is paramount [2].

Polycystic Ovary Syndrome (PCOS), a widespread endocrine disorder, poses significant diagnostic challenges due to its complex symptom profile and associated metabolic risks. Using clinical and lifestyle data, this study evaluates the predictive capabilities of seven machine learning models for PCOS classification. Logistic Regression (LR) emerges as the most effective algorithm, achieving the highest scores in accuracy (91.7%), precision (96%), and ROC AUC (96.8%). The superior performance of LR is enhanced through the use of Synthetic Minority Over-sampling Technique (SMOTE) for addressing class imbalance and ANOVA F-score feature selection for identifying key predictors. The model's interpretability and simplicity position it as a practical solution for clinical decision-support systems, facilitating early diagnosis and intervention while maintaining transparency in healthcare settings [3].

These studies collectively underscore the transformative potential of data-driven methodologies in addressing real-world issues with accuracy, speed, and adaptability. Whether through predictive environmental analytics, refined numerical modeling, or intelligent healthcare diagnostics, the featured research reaffirms the indispensable role of computational science in advancing societal resilience and technological progress.

References:

[1] R. Taha, F. Musleh, A. Rahman Musleh, "Fire Type Classification in the USA Using Supervised Machine Learning Techniques," *Journal of Engineering Research and Sciences*, vol. 4, no. 6, pp. 1–8, 2025, doi:10.55708/js0406001.

[2] Dalabaev Umurdin, Khasanova Dilfuza, "Analysis of Difference Schemes of Two-Point Boundary Value Problems using the Method of Moving Nodes," *Journal of Engineering Research and Sciences*, vol. 4, no. 6, pp. 9–15, 2025.

[3] R. Taha, H. Zain El Abdin, T. Musleh, "Comparative Analysis of Supervised Machine Learning Models for PCOS Prediction Using Clinical Data," *Journal of Engineering Research and Sciences*, vol. 4, no. 06, pp. 16–26, 2025, doi:10.55708/js0406003.

Editor-in-chief

Dr. Jinhua Xiao

JOURNAL OF ENGINEERING RESEARCH AND SCIENCES

Volume 4 Issue 6 **June 2025**

CONTENTS

Received: 13 January 2025, Revised: 07 May 2025, Accepted: 15 May 2025, online: 16 June 2025

DOI: https://doi.org/10.55708/js0406001

Fire Type Classification in the USA Using Supervised Machine Learning Techniques

Ranyah Taha*[,1], Fuad Musleh[2], Abdel Rahman Musleh[3]

[1] Computer Science Dept., Al-Iman School, Bahrain

[2] Civil engineering Department, College of Engineering, University of Bahrain, Sakhir, 1054, Bahrain

[3] Electrical and Electronics Engineering Department, College of Engineering, University of Bahrain, Sakhir, 1054, Bahrain

*Corresponding author: Ranyah Taha, raniacs2014@gmail.com

ABSTRACT: Wildfires are a growing global concern, causing widespread environmental, economic, and health impacts. In the USA, fire incidents have become more frequent and intense due to factors such as climate change, prolonged droughts, and human activities. Machine learning plays a vital role in predicting and classifying fires by analyzing vast satellite and environmental datasets with high speed and accuracy. These models support early warning systems and informed decision-making, ultimately helping to reduce damage and improve emergency response strategies. This study evaluates the effectiveness of supervised machine learning algorithms—including Decision Tree (DT), Random Forest (RF), Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Gradient Boosting Classifier (GBC)—in classifying different fire types. The DT emerges as the top-performing model, achieving the highest results across all evaluation metrics, including 96.69% accuracy, precision, recall, and F1 score. RF follows closely with similarly strong performance, making it a highly reliable alternative. GBC ranks third, showing balanced and consistent results above 92% in all metrics. In contrast, SVC and LR perform less effectively, particularly in precision and F1 score, indicating that they are not ideal choices for fire type classification in this study. The novelty of this study lies in its application of a comparative ML framework to classify fire types using real satellite-based observations specific to the USA. region. By integrating and evaluating multiple ML models on this large-scale, real-world dataset, the study provides valuable insights into model suitability for fire classification tasks and offers practical guidance for deploying predictive tools in environmental monitoring and disaster management systems.

KEYWORDS: Artificial Intelligence, Data Analysis, Fire type Classification, Machine Learning, USA, NASA, Civil Engineering.

1. Introduction

Fires represent a major environmental disaster due to their rapid spread, the complexity of containment efforts, and the extensive damage they inflict on ecosystems, infrastructure, and human health. In the USA, fire incidents—particularly wildfires—have become increasingly frequent and intense, driven by factors such as climate variability, land use changes, and human activity. The severe consequences of these events have underscored the importance of fire detection, classification, and management, making fire monitoring a vital component of forestry, environmental protection, and emergency response strategies [1].

Several critical factors contribute to the occurrence and spread of fires across the United States. Climatic variables—including high temperatures, strong wind speeds, low relative humidity, limited rainfall, and lightning probability—create conditions that significantly increase the risk of fire ignition and propagation. In addition to environmental influences, human-related factors such as population density, land development, and increased recreational or industrial activity in forested and rural regions further elevate fire risk. The combination of these natural and anthropogenic elements makes fire prediction and classification an increasingly urgent priority for disaster management and environmental protection [1].

Artificial Intelligence (AI) plays a transformative role in modern wildfire detection and classification systems, significantly enhancing the ability to anticipate, monitor,

and manage fire events. AI technologies contribute to various aspects of wildfire preparedness and response, including fuel assessment, fire behavior prediction, real-time detection, impact estimation, and strategic fire management. Leveraging tools such as satellite imagery, historical weather data, and computational models, AI enables the automated analysis of complex environmental patterns [2].

In particular, Machine Learning (ML)—a subset of AI—is increasingly utilized for the early prediction and accurate classification of fires by identifying patterns in large-scale datasets. These intelligent systems support timely decision-making and resource allocation, making AI a critical component in reducing wildfire-related risks and improving emergency response strategies [2].

This study utilizes a dataset comprising fire incident records detected throughout the United States in 2021. The data were collected by the VIIRS sensor aboard the SNPP satellite and sourced from the NASA Open Data Portal. The research follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to ensure a structured approach to data analysis and model development. Since each machine learning method has its own advantages and limitations, a comparative evaluation is necessary to determine the most effective model for classifying fire types. Therefore, this work focuses on assessing the performance of six supervised learning algorithms—DT, RF, SVC, KNN, LR, and GBC—in predicting fire categories. The paper is organized into several sections: a literature review, methodology, data description and preprocessing, model implementation, results, discussion, conclusion, and future recommendations.

2. Literature Review

Several ML algorithms have been instrumental in advancing forest fire forecasting. This section reviews various studies that have applied these methods, as outlined below recent research has extensively explored various ML and AI techniques for forest fire prediction and management.

In [3], the authors addressed critical challenges in forest fire prediction by proposing a robust ML framework specifically designed to handle severely imbalanced datasets, a frequent issue in wildfire modeling. The study utilized Copernicus reanalysis data from 2000 to 2018, incorporating 27 features including temperature, soil moisture, wind speed, and vegetation indices to model fire susceptibility in Canada's boreal forests. To manage the 158:1 non-fire-to-fire ratio, the authors employed a hybrid sampling strategy combining NearMiss3 for undersampling and SMOTE-ENN for oversampling with noise reduction. Among the models tested—RF, XGB, LGBM, and CatBoost—XGB combined with NearMiss3 at a 0.09 sampling ratio achieved optimal performance, with 98.08% accuracy, 86.06% sensitivity, and 93.03% specificity. Moreover, the study emphasized the balance between computational efficiency—demonstrated by LGBM's histogram-based learning—and model interpretability, using feature importance to highlight soil moisture as a dominant factor in fire prediction.

Similarly, the authors in [4] conducted a detailed evaluation of ML models using meteorological data from Algeria, integrating a temporal-stage approach and correlation-based feature selection (CFS) to enhance predictive accuracy. The study divided the dataset into six-time intervals and focused on weather indicators such as temperature, humidity, and FWI components. Important predictors including FFMC, DMC, and FWI were identified through CFS, significantly improving model accuracy. Among the tested models—DT, RF, SVC, LR, KNN, and GNB—DT and RF both achieved perfect accuracy (100%) during the peak fire season (June–July), outperforming SVC, LR, and KNN, each of which recorded 98%. The authors also observed that variables like wind speed contributed minimally, reinforcing the need for region-specific features in fire prediction. Although GBC was not part of the study, the findings strongly support the use of ensemble and tree-based methods for regionally adapted fire forecasting, particularly within U.S. contexts.

In another effort to improve prediction through model integration, the authors in [5] employed an ensemble-based soft voting strategy combining DT, KNN, and LR to map wildfire susceptibility in Iran's Alborz Mountains. Using MODIS thermal anomaly data and a GPS-corrected fire inventory, the study incorporated 17 variables across anthropogenic, vegetation, topographic, climatic, and hydrological domains. The ensemble model achieved an average AUC of 88%, peaking at 93% in one-fold during 10-fold cross-validation, surpassing the performance of each individual base classifier. The generated susceptibility map classified the landscape into five risk zones, revealing that 21% of the area was at high or very high risk—correlating well with historical fire records. The study underscored the benefits of ensemble learning for improving accuracy and robustness, and suggested that integrating more advanced models like RF or GBC into such frameworks could further improve adaptability across diverse USA terrains.

Expanding the geographical scope, the authors in [6] conducted a large-scale comparative study involving more than 1.04 million fire events from the USA (1992–2015) and 517 cases from Portugal (2000–2003). The dataset featured a wide range of spatial, temporal, and environmental variables. A variety of models—DL, DT, SGD, ExGBT, and LR—were evaluated for wildfire size

classification, with results showing accuracy ranging from 80% to 82%. DT and ExGBT outperformed others, while GA was employed to derive symbolic representations of wildfire behavior, producing correlation coefficients above 0.80. To enhance balance and interpretability, SMOTE was used to address class imbalance, and SHAP values revealed temperature and weather indices as critical predictive factors. The study demonstrates the value of combining performance-focused models with interpretable AI techniques, especially when handling large, complex wildfire datasets like those found in the U.S.

On a global scale, in [7], the authors used high-resolution (0.25°) global data from 2015 to evaluate wildfire susceptibility based on meteorological variables, fire weather indices, and anthropogenic influences. Models assessed included RF, XGB, and MLP, benchmarked against traditional LR and linear regression. The XGB model yielded the highest performance with an AUC of 97% for wildfire occurrence and a MAE of 3.13 km² for burned area prediction. SMOTE and class-weighted loss functions were used to mitigate data imbalance, while SHAP analysis identified key variables such as historical fire activity, relative humidity, and precipitation.

Although the study aimed for global applicability, regional analysis showed that ML models performed better in Africa and Asia, while in North America, traditional fire indices remained relevant. These findings reinforce the effectiveness of ensemble and deep learning models like XGB and MLP, particularly in high-dimensional, data-rich environments such as the U.S.

In the context of localized prediction, in [8], the authors applied several ML models to Greece's Attica basin, using a custom dataset with 12 meteorological features including temperature, humidity, wind, and rainfall. The study explored binary classification (fire/no fire), multiclass classification (fire severity), and regression (burned area prediction). Among the tested models—RF, XGB, KNN, NN, SVM, LR, and DT—RF performed best for binary classification with 70% accuracy using all features, XGB was most effective with a reduced four-feature set (67.4% accuracy), and KNN achieved the highest R^2 score of 70% for regression. Validation against the Montesinho dataset supported the generalizability of the approach, suggesting its adaptability to fire-prone regions in the USA.

Similarly, the authors in [9] proposed an ML-driven prediction framework utilizing meteorological variables and FWI data from Portugal's Montesinho Park. The study tested RF, SVM, GBC, LR, and K-means, using stepwise regression and backward elimination for feature selection. Temperature and humidity were identified as the most influential features. SVM and RF performed best in estimating burned areas. While regression performance was modest (R^2 = 14%), clustering via K-means (optimized with the elbow method) allowed for localized fire risk assessment. The authors emphasized the value of incorporating spatial and climatic diversity into prediction models—especially relevant to U.S. regions like California and the Pacific Northwest—and suggested further improvements including vegetation types, forest density, and ignition source modeling.

Building on the comparison of classifiers, in [10], the authors evaluated the performance of RF, SVM, DT, and NB and identified RF as the most accurate model for wildfire forecasting. Their findings highlight RF's reliability in supporting early warning and fire response efforts. Similarly, in [11], the authors affirmed RF as the top-performing algorithm among the same set, emphasizing its critical role in risk reduction strategies.

The reviewed literature reflects the increasing reliance on advanced ML techniques for wildfire prediction and classification, particularly ensemble and tree-based models such as RF, XGB, LGBM, CatBoost, DT, GBC, and AdaBoost. These models consistently outperform traditional approaches like LR and linear regression, especially when combined with strategies such as SMOTE, correlation-based and stepwise feature selection, and SHAP for model interpretability. Other algorithms including SVM, KNN, GNB, SGD, MLP, NN, and GA have also demonstrated strong performance in specific tasks, such as burned area regression and symbolic modeling. Unsupervised methods like K-means have been effectively used for spatial clustering and localized risk assessment. The studies emphasize the importance of regional and temporal adaptation, the integration of spatial and environmental data, and handling class imbalance. Although challenges remain in accurately modeling fire extent, ensemble and hybrid methods show strong potential. Overall, the literature confirms the adaptability and scalability of a wide array of ML models for wildfire forecasting across the diverse climatic zones of the U.S.

3. Research Methodology and approach

3.1. Background of the Research Study

This research was conducted using the Google Collab platform as the primary workspace, with Scikit-learn serving as the main Python library for implementing machine learning models. A total of six algorithms—DT, RF, SVC, KNN, LR, and GBC—were employed to explore and analyze the dataset. The study adopted the CRISP-DM methodology, a widely accepted framework for machine learning projects. This methodology comprises six essential phases: identifying the project goals (business understanding), examining the dataset (data understanding), preparing the data for analysis (data

preparation), building and optimizing models (modeling), evaluating the performance of those models (evaluation), and making the model ready for real-world use (deployment) [2]. Utilizing this structured approach ensured clarity and efficiency throughout the process, contributing to the reliable and accurate results illustrated in Figure 1.

Figure 1: Phases of the CRISP-DM Methodology.

3.2. Dataset Description

The data set used in this study contains records of fire incidents detected across the USA during the year 2021. These observations were captured by the VIIRS sensor on board the Suomi National Polar-orbiting Partnership (SNPP) satellite and obtained through the NASA Open Data Portal [12]. This open-access platform provides researchers with dependable, high-resolution datasets crucial for advancing studies in renewable energy and enhancing grid management strategies. It delivers comprehensive information on solar radiation, meteorological variables, and atmospheric conditions, which are instrumental in building precise energy forecasting models and tackling the unpredictability inherent in renewable energy systems. Furthermore, the platform supports sophisticated simulations and machine learning applications, contributing to more accurate predictive analytics and improved grid efficiency. Its commitment to open data access fosters cross-disciplinary research and innovation, establishing it as a vital resource for environmental and energy research communities [12].

The dataset includes 661,058 records, comprising 360,993 nighttime and 300,065 daytime entries. It features eight input variables and one categorical target variable, which classifies fire events into four categories: Type 0 (presumed vegetation fires), Type 1 (active volcanic activity), Type 2 (fires from stationary land-based sources), and Type 3 (offshore fire detections over water bodies).

This classification framework underscores the dataset's emphasis on distinguishing between different fire origins and behaviors [12]. A summary of the dataset's attributes is provided in Table 1.

Table 1: Dataset Description

Attribute	Definition	Datatypes
Bright_ti4	Measures the brightness temperature in Band 4 of the thermal infrared spectrum (TIR).	Float64
Bright_ti5	Measures the brightness temperature in Band 5 of the TIR.	Float64
Scan	Measures the satellite's scanning ability, including angle, direction, and spatial coverage.	Float64
Track	Describes the satellite's orbital path, alongside its current location and trajectory.	Float64
FRP	Fire radiative power (MW).	Float64
Latitude	Fire pixel latitude(degree).	Float64
Longitude	Fire pixel longitude (degree).	Float64
Day-night	Uses the solar zenith angle (SZA) to determine whether conditions are day or night.	Object
Type	Type attributed to thermal anomaly.	Object

3.3. Dataset Preparation

Following the data exploration phase, the preparation of the dataset is initiated. This stage involves multiple preprocessing steps, including managing missing values, removing duplicate entries, applying normalization techniques, selecting relevant features, encoding categorical variables, and dividing the data into training and testing sets. These steps are essential to ensure the dataset is clean, structured, and ready for effective modeling and further analytical procedures.

3.3.1. Missing Data

To verify the integrity of the dataset, two standard functions were employed: isnull().sum() and duplicated().sum() [13]. The isnull(). sum() function is used to detect and count any missing values across the dataset columns, while duplicated().sum() identifies repeated rows that could compromise data quality. The execution of these checks revealed that the dataset contained neither missing values nor duplicate entries. This confirmation of data completeness and consistency contributes to improved data quality, which is critical for building accurate and reliable machine learning models.

3.3.2. Balancing the Dataset

The distribution of fire types in the dataset reveals a significant imbalance, with Type 0 (presumed vegetation fires) dominating at 86.88% of the total records. In contrast, the other categories are considerably less represented, especially Type 1 (active volcano), which constitutes only 0.10%. To address this disparity and enhance the performance of machine learning models across all classes, the dataset was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) technique prior to training. SMOTE is a popular technique used in imbalanced classification problems to

help balance the dataset by generating synthetic data points for the minority class [14].

3.3.3. *Encoding Categorical Data*

The dataset underwent label encoding to transform categorical variables into numeric format, an essential preprocessing step since most machine learning algorithms require numerical input [15].

In this study, fire incidents were categorized according to their type: Type 0 representing presumed vegetation fires, Type 1 indicating volcanic activity, Type 2 referring to stationary land-based fires, and Type 3 covering offshore fire detections over water. This conversion was vital to ensure the data was compatible with the classification models, thereby improving the effectiveness and accuracy of the training process.

3.3.4. *Splitting Data*

Initially, the dataset was split into two parts: 80% for training and 20% for testing. This division allows the model to learn from the majority of the data while reserving a portion for evaluating its performance on unseen examples.

3.3.5. *Data Normalization*

The numerical features bright_ti4, bright_ti5, scan, track, and frp were normalized to bring their values within a consistent range, such as 0 to 1 or -1 to 1 [16]. This scaling process ensures that each feature contributes equally during model training, preventing any one variable from disproportionately influencing the learning process and supporting more balanced, unbiased model performance.

3.4. *Modelling*

Six machine learning algorithms—DT, RF, SVC, KNN, LR, and GBC—were implemented to classify the fire types.

Decision Tree (DT) is a non-parametric learning method that uses a tree-like structure to make decisions based on feature thresholds. It recursively splits the dataset into subsets based on the most significant feature at each node, making it interpretable and efficient for handling both categorical and numerical data. However, it is prone to overfitting, particularly on noisy datasets [15].

Random Forest (RF) is an ensemble learning technique that builds multiple decision trees during training and merges their outputs for improved accuracy and robustness. By averaging the results (in classification, via majority voting), RF reduces overfitting and variance compared to individual trees, offering better generalization on unseen data [15].

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm that classifies data points based on the majority label among their k-nearest neighbors in the feature space. Though computationally intensive during prediction, KNN is intuitive and works well with non-linear data distributions when appropriate distance metrics and normalization are applied [17].

Logistic Regression (LR) is a statistical model that uses the logistic function to model the probability of a binary or multiclass outcome. Despite its simplicity, LR is a strong baseline model due to its efficiency, interpretability, and solid performance in linearly separable problems [18].

Gradient Boosting Classifier (GBC) is a powerful ensemble method that builds models sequentially, where each new model attempts to correct the errors made by the previous ones. It combines weak learners (typically shallow trees) using gradient descent optimization to minimize the loss function, achieving high predictive accuracy at the cost of increased training time [16].

Support Vector Classifier (SVC) is based on the principles of Support Vector Machines (SVM). It attempts to find the optimal hyperplane that best separates the data into distinct classes by maximizing the margin between support vectors. SVC is especially effective in high-dimensional spaces and is robust to overfitting when the kernel and regularization parameters are properly selected [19] .

3.5. *Performance Evaluation*

The effectiveness of the supervised machine learning models is evaluated using key performance metrics, including accuracy, recall, F-measure and precision, which collectively provide insight into their classification performance.

3.5.1. *Accuracy*

It represents the proportion of correctly predicted instances out of the total number of predictions made. It reflects the overall effectiveness of a model in classifying both positive and negative cases correctly shown in equation (1) [15].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

3.5.2. *F-measure*

It offers a balanced assessment by combining both metrics into a single value, especially useful when the data is imbalanced or when equal consideration of false positives and false negatives is needed shown in equation (2) [15].

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (2)$$

3.5.3. *Precision*

It measures the ratio of correctly predicted positive instances to the total predicted positives. It indicates how many of the instances labeled as positive by the model are actually relevant, helping to evaluate the model's reliability in making positive predictions shown in equation (3) [17].

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

3.5.4. *Recall*

It refers to the proportion of actual positive cases that are correctly identified by the model. It is particularly important in situations where missing positive cases is costly or undesirable shown in equation (4) [18].

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

4. Results

In terms of accuracy, DT attains the top performance with 96.69%, closely followed by RF at 96.37%, both demonstrating strong capabilities in correctly identifying fire types. GBC also delivers notable accuracy at 93.16%, with KNN achieving 91.27%. On the other hand, SVC and LR register comparatively lower accuracy rates of 88.35% and 87.58%, respectively, suggesting relatively less effective classification results, as illustrated in Table 2 and Figure 2.

Looking at precision, DT again leads with 96.70%, indicating a high level of accuracy in its positive predictions and a minimal rate of false positives. RF follows closely with a precision of 96.31%, while GBC achieves 92.76%, both reflecting reliable classification outputs. KNN also shows solid results with 90.57%, whereas SVC and LR lag behind at 83.61% and 83.65%, respectively, highlighting a greater occurrence of incorrect positive classifications.

Regarding recall, which assesses the ability to correctly identify actual fire instances, DT maintains its lead at 96.69%, with RF slightly behind at 96.37%. GBC continues to perform well with 93.16%, while KNN records 91.27%. In contrast, SVC and LR exhibit lower recall rates of 88.35% and 87.58%, indicating a higher chance of failing to detect true fire occurrences.

When considering the F1 score, which harmonizes precision and recall into a single performance metric, DT secures the highest value at 96.67%, confirming its balanced and robust classification ability. RF follows with an F1 score of 96.19%, and GBC reaches 92.67%. KNN also maintains dependable performance with 90.79%. Meanwhile, SVC and LR yield lower F1 scores of 85.50% and 84.71%, respectively, indicating limitations in managing the trade-off between precision and recall.

Table 2: Performance Comparison between models.

Model	Accuracy (%)	Recall (%)	Presion (%)	F1-Scor (%)
SVC	88.35	88.35	83.61	85.50
RF	96.37	96.37	96.31	96.19
KNN	91.27	91.27	90.57	90.79
LR	87.58	87.58	83.65	84.71
DTC	96.69	96.69	96.70	96.70
GBC	93.16	93.16	92.76	92.67

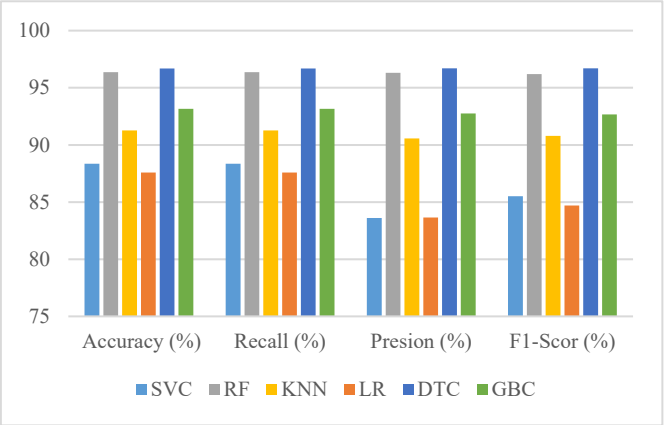

Figure 2: Performance Plot of Proposed Models

5. Discussion

The findings of the current study, which evaluates six supervised ML models—DT, RF, GBC, KNN, SVC, and LR—for fire type classification, align well with trends observed in the reviewed literature while also offering noteworthy advancements in model performance and application specificity.

In this study, DT achieved the highest accuracy (96.69%), precision (96.70%), recall (96.69%), and F1 score (96.67%), outperforming other models. These results are consistent with the findings of Khosravi et al., who reported perfect classification accuracy for DT and RF during peak fire seasons in Algeria, confirming the effectiveness of tree-based models in wildfire classification tasks. Similarly, RF performed robustly across all metrics in the current study—attaining 96.37% accuracy and 96.19% F1 score—which echoes its dominant position in several previous studies, including those by Tavakoli, Barzani et al., and Al-Bashiti & Naser, where RF either matched or exceeded other ensemble models in terms of predictive accuracy and interpretability.

GBC also demonstrated strong performance in this work, with consistent results across accuracy (93.16%), precision (92.76%), recall (93.16%), and F1 score (92.67%). While GBC was not explicitly evaluated in some past works such as those by Khosravi et al., its potential was highlighted in Chaubey et al. and Alkhatib et al., who supported the integration of ensemble models to improve classification reliability—particularly when using complex and high-dimensional environmental data.

KNN, although not an ensemble method, delivered solid results (accuracy: 91.27%, F1 score: 90.79%), which aligns with Stafylas Demetrios' regression-based analysis, where KNN showed competitive performance in predicting burned area. However, KNN remains sensitive to feature scaling and may not capture complex decision boundaries as effectively as tree-based models, which is reflected in its slightly lower scores compared to DT, RF, and GBC.In contrast, SVC and LR showed the weakest performance across all metrics. SVC recorded 88.35% accuracy and 85.50% F1 score, while LR followed closely behind with 87.58% accuracy and 84.71% F1 score. These outcomes are consistent with earlier studies, such as those by Al-Bashiti and Naser, where LR underperformed relative to ensemble and tree-based models, and by Shmuel and Heifetz, who showed that while traditional models like LR offer baseline predictability, they fall short in handling the nonlinear and complex nature of wildfire dynamics.

Another important point of comparison is how well the current study addresses model balance. Unlike some previous works that focused on peak fire seasons or lacked formal imbalance-handling strategies, this study ensured an equal class distribution prior to training, which likely contributed to the high and consistent scores for DT, RF, and GBC across all evaluation metrics. This balanced approach strengthens the reliability and generalizability of the findings, especially for real-world applications in USA fire forecasting, where underrepresented classes often challenge prediction accuracy.

Furthermore, this study's comparative framework adds value by using a unified dataset and standardized preprocessing, enabling a fair and direct performance comparison. While prior literature often evaluated models on region-specific or task-specific datasets (e.g., ignition, size, burned area), this study provides a focused comparison on fire type classification, offering insights particularly useful for U.S.-based fire management systems aiming for categorical fire event identification.

6. Conclusion and Future Directions

This study assessed the effectiveness of six supervised machine learning algorithms—DT, RF, GBC, KNN, SVC, and LR—in classifying fire types in the United States using satellite-derived data. Among the evaluated models, DT consistently achieved the best results, recording the highest scores in accuracy (96.69%), precision (96.70%), recall (96.69%), and F1 score (96.67%). RF closely followed, while GBC also demonstrated strong and balanced performance across all metrics. In contrast, SVC and LR exhibited comparatively lower predictive capabilities, highlighting their limitations in capturing the complex, nonlinear patterns characteristic of fire behavior.

These findings align with previous research, where tree-based and ensemble models—particularly DT, RF, and XGB—have repeatedly proven effective in wildfire prediction. Their success can be attributed to several key strengths. First, these models are well-suited to capturing nonlinear interactions among environmental variables such as temperature, humidity, wind, and vegetation, which are critical in fire dynamics. Second, they effectively manage heterogeneous and high-dimensional datasets, including those combining meteorological indices, satellite imagery, and geospatial information. Third, they demonstrate robustness to noise, missing values, and outliers, enabling more reliable predictions in real-world conditions.

Moreover, ensemble methods such as RF and XGB offer enhanced generalization through the aggregation of multiple decision paths, thereby reducing the risk of overfitting. These models also support model interpretability through feature importance rankings and SHAP analysis, providing valuable insights into the most influential factors driving fire classifications—an essential feature for transparent and accountable decision-making in wildfire management systems.

By applying a balanced dataset and a standardized evaluation framework, this study provides a robust comparison of model performance, contributing novel insights to the evolving field of ML-driven wildfire forecasting. The findings reaffirm that tree-based and ensemble algorithms are not only highly accurate but also scalable, flexible, and interpretable, making them particularly well-suited for operational deployment in real-world fire risk management applications—especially across the diverse climatic and ecological regions of the USA.

Looking forward, future research should explore the integration of real-time meteorological feeds, higher-resolution spatial data, and advanced ensemble strategies such as model stacking and hybrid architectures. Additionally, incorporating deep learning techniques and spatiotemporal modeling could further enhance predictive precision, enabling more dynamic and proactive wildfire forecasting systems capable of addressing both localized threats and broader regional patterns.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The Authors hereby acknowledge that the funding of this paperwork was done and shared across all Authors concerned.

References

[1] A. Modaresi Rad et al., "Human and infrastructure exposure to large wildfires in the United States," Nature Sustainability, vol. 6, no. 11, pp. 1343-1351, 2023, doi:10.1038/s41893-023-01163-z.

[2] S. P. H. Boroujeni et al., "A comprehensive survey of research towards AI-enabled unmanned aerial systems in pre-, active-, and post-wildfire management," Information Fusion, p. 102369, 2024, doi:org/10.1016/j.inffus.2024.102369.

[3] F. Tavakoli, "Dataset Creation and Imbalance Mitigation in Big Data: Enhancing Machine Learning Models for Forest Fire Prediction," University of Waterloo, 2023, http://hdl.handle.net/10012/20046.

[4] H. Khosravi, M. R. Shafie, A. S. Raihan, M. A. B. Syed, and I. Ahmed, "Optimizing Forest Fire Prediction: A Comparative Analysis of machine learning models through feature selection and time-stage evaluation," Preprints. org, 2023, doi: 10.20944/preprints202312.0577.v1

[5] A. Rezaei Barzani, P. Pahlavani, and O. Ghorbanzadeh, "Ensembling of decision trees, KNN, and logistic regression with soft-voting method for wildfire susceptibility mapping," ISPRS Annals of the Photogrammetry, Remote Sensing Spatial Information Sciences, vol. 10, pp. 647-652, 2023, doi:10.5194/isprs-annals-X-4-W1-2022-647-2023, 2023.

[6] M. K. Al-Bashiti and M. Naser, "Machine learning for wildfire classification: Exploring blackbox, eXplainable, symbolic, and SMOTE methods," Natural Hazards Research, vol. 2, no. 3, pp. 154-165, 2022, doi:10.1016/j.nhres.2022.08.001.

[7] A. Shmuel and E. Heifetz, "Global wildfire susceptibility mapping based on machine learning models," Forests, vol. 13, no. 7, p. 1050, 2022, doi:10.3390/f13071050.

[8] D. Stafylas, "Wildfire prediction using machine learning," M.S. thesis, University of West Attica, 2022.

[9] T. Preeti, S. Kanakaraddi, A. Beelagi, S. Malagi, and A. Sudi, "Forest fire prediction using machine learning techniques," in 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-6: IEEE, DOI:10.1109/CONIT51480.2021.9498448.

[10] R. Alkhatib, W. Sahwan, A. Alkhatieb, and B. Schütt, "A brief review of machine learning algorithms in forest fires science," Applied Sciences, vol. 13, no. 14, p. 8275, 2023, doi:10.3390/app13148275.

[11] F. N. Ismail, B. J. Woodford, S. A. Licorish, and A. D. Miller, "An assessment of existing wildfire danger indices in comparison to one-class machine learning models," Natural Hazards, pp. 1-32, 2024, doi:10.1007/s11069-024-06738-3.

[12] " NASA Open Data Portal https://data.nasa.gov/browse " 2021.

[13] S. Alshakrani, R. Taha, and N. Hewahi, "Chronic kidney disease classification using machine learning classifiers," in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021, pp. 516-519: IEEE, doi: 10.1109/3ICT53449.2021.9581345.

[14] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem: A review," in 2021 sixth international conference on informatics and computing (ICIC), 2021, pp. 1-8: IEEE, doi: 10.1109/ICIC54025.2021.9632912.

[15] F. A. Musleh and R. G. Taha, "Forecasting of forest fires using machine learning techniques: a comparative study," in 6th Smart Cities Symposium (SCS 2022), 2022, vol. 2022, pp. 337-342: IET, doi: 10.1049/icp.2023.0571.

[16] F. Ahmad Musleh, "A Comprehensive Comparative Study of Machine Learning Algorithms for Water Potability Classification," International Journal of Computing Digital Systems, vol. 15, no. 1, pp. 1189-1200, 2024, doi: 10.1049/icp.2023.0571.

[17] R. Taha, S. Alshakrani, and N. Hewahi, "Exploring Machine Learning Classifiers for Medical Datasets," in 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 255-259: IEEE, doi: 10.1109/ICDABI53623.2021.9655862.

[18] A. M. Zeki, R. Taha, and S. Alshakrani, "Developing a predictive model for diabetes using data mining techniques," in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021, pp. 24-28: IEEE, doi: 10.1109/3ICT53449.2021.9582114.

[19] F. A. Musleh, "A comparative study to forecast the total nitrogen effluent concentration in a wastewater treatment plant using machine learning techniques," International Journal of Computing Digital Systems, vol. 14, no. 1, pp. 10447-10456, 2023.

Mrs. Ranyah Taha completed her MSc in Big Data Science and Analytics in 2022 through a joint program between Liverpool John Moores University and the University of Bahrain. She earned her BSc in Computer Science from the University of Bahrain in 2018. Her research focuses on leveraging Data Science and Analytics, particularly Machine Learning and Deep Learning, to build advanced models and extract valuable insights from complex datasets. She has contributed to many research papers and was awarded the NASA International Space Apps Challenge – Space Apps Bahrain 2023 Local Impact Award.

Dr. Fuad Musleh has received his Ph.D. and M.Sc. degrees from the University of Alabama in Huntsville, and his B.Sc. from Jordan University of Science and Technology in Jordan. He is currently serving as an Assistant Professor at the University of Bahrain. His research interests include environmental and water resource engineering, particularly in vegetation–flow interaction and environmental conservation. He also focuses on data analysis applications in environmental systems and sustainability science.

Mr. Abdel Rahman Musleh is currently a senior undergraduate student pursuing a B.Sc. in Electrical Engineering at the University of Bahrain. His academic focus lies in developing cyber-physical systems that integrate artificial intelligence (AI) and machine learning (ML) to enhance system efficiency, reliability, and sustainability.His interests include the application of ML in renewable energy systems and smart grid infrastructure, with a growing involvement in research related to intelligent automation and real-time simulations.

Received: 24 April, 2025, Revised: 22 May, 2025, Accepted: 23 May, 2025, Online: 22 June, 2025

DOI: https://doi.org/10.55708/js0406002

Analysis of Difference Schemes of Two-Point Boundary Value Problems using the Method of Moving Nodes

Dalabaev Umurdin *, **Khasanova Dilfuza**

University of World Economy and Diplomacy, Tashkent, Uzbekistan

*Corresponding author: Dalabaev Umurdin, Buyuk Ipak Yuli 54, +998910091309, udalabaev@uwed.uz

ABSTRACT: This article addresses the calculation of approximation errors in numerical methods for solving differential equations. A fundamental challenge when replacing differential equations with discrete representations is ensuring that the discrete solution closely approximates the exact solution. To tackle this, a grid area is established for the difference solution, with discrete solutions evaluated at specific nodal points. Traditionally, the degree of approximation in this context is expressed using the notation $O(h^p)$, where h represents the grid step and p indicates the order of accuracy. A significant advancement in this area is the application of the moving nodes method, which enables the calculation of approximation errors at these nodal points. This method allows researchers to derive an approximate analytical expression for the discrete solution, which serves as a foundation for calculating the approximation error.

KEYWORDS: Moving Node Method, Approximation error, To-Point Boundary Problem

1. Introduction

This article is an expanded version of the article presented in [1]. The numerical solution methods for differential equations fundamentally rely on transforming differential problems into difference problems [2–5]. In simpler terms, solving differential equations requires understanding how to approximate them. This involves converting a differential equation into a system of algebraic equations, which is based on the values of the desired functions at specific points on a grid. Recent studies [6]–[11] have introduced a new approach for approximating differential operators, enhancing the accuracy and efficiency of these methods. One of the significant advantages of the moved node method is that it enables the calculation of an explicit expression for the approximation error when replacing differential equations with difference ones. Understanding this error is crucial because it provides insights into the reliability and accuracy of the numerical solution. By quantifying the error, researchers can refine their methods and improve the overall quality of the numerical solutions obtained.

In conclusion, the transformation of differential equations into difference equations is a fundamental aspect of numerical analysis. The development of innovative methods like the moved node method represents a significant advancement in this field, providing researchers and practitioners with powerful tools to tackle complex differential problems more effectively. As numerical methods continue to evolve, the importance of understanding and minimizing approximation errors will remain a critical area of focus for ensuring the accuracy and reliability of solutions.

On the basis of the movable node, an approximate analytical expression for the difference solution of the differential problem was obtained [12]. This development represents a significant step forward in numerical methods, as it provides a more refined approach to approximating solutions to differential equations. The analytical expression derived from the movable node approach allows for greater flexibility and accuracy when dealing with complex differential problems.

In [13], the moving nodes method was further applied to construct the control volume method, which is widely used in computational fluid dynamics and other engineering applications.

In [14], the authors explored the potential to increase accuracy by combining the moving nodes method with the ideas of Richardson's extrapolation. Richardson's extrapolation is a technique used to improve the precision of numerical approximations by utilizing solutions obtained at different grid resolutions. By integrating this method with the moving nodes approach, it is possible to achieve higher-order accuracy in the numerical solutions, thereby reducing the error associated with the approximation.

Some questions regarding the monotonicity of the difference scheme using the movable node are addressed in [15]. Monotonicity is an important property in numerical methods, as it ensures that the numerical solution behaves in a physically realistic manner, avoiding non-physical oscillations or spurious solutions. Understanding and ensuring the monotonicity of the difference scheme is crucial for maintaining the stability and reliability of the numerical method, especially in problems involving sharp gradients or discontinuities.

The application of the moving nodes method to various applied problems is reflected in [16]. This demonstrates the versatility of the method across different fields, such as fluid dynamics, heat transfer, and structural analysis.

Moreover, based on the choice of the velocity profile on the edge of the control volume, qualitative schemes were obtained in [17]. The velocity profile plays a critical role in determining the flow characteristics and behavior within the control volume.

In summary, the integration of the movable node method into various numerical frameworks and its application to real-world problems highlights its significance in advancing numerical analysis. The ongoing exploration of its properties, such as accuracy, monotonicity, and adaptability to different contexts, continues to enhance the capability of numerical methods in solving complex differential equations effectively. As research in this area progresses, the potential for further innovations and improvements remains substantial, promising even greater advancements in the field of numerical solutions.

This paper describes the application of the moving nodes method to the calculation of the approximation error. The moving nodes method provides a dynamic approach to numerical analysis, allowing for the adjustment of grid points based on the behavior of the solution.

When a two-point boundary value problem is solved using difference methods, the question of the degree of approximation typically arises. This degree of approximation is crucial as it directly impacts how closely the numerical solution aligns with the exact solution. In numerical analysis, understanding the closeness of the exact solution to its approximation is essential for evaluating the effectiveness of the chosen method.

The quality of the difference scheme is often assessed based on this degree of approximation. A higher degree indicates a more accurate representation of the solution, while a lower degree suggests potential discrepancies that may arise from the numerical method employed. This evaluation is typically conducted by analyzing the behavior of the approximation error, which quantifies the difference between the exact solution and the numerical approximation.

Interestingly, in this analysis, other parameters—such as the coefficients of the differential equation—are not explicitly involved in the expression for the approximation error. This is significant because it allows researchers to focus on the fundamental aspects of the numerical method without being distracted by the specific characteristics of the differential equation being solved. By isolating the approximation error from these coefficients, the analysis can yield more generalized insights into the behavior of the numerical solution.

Obtaining an explicit expression allows researchers to identify how changes in the grid size, the choice of the moving nodes, and other factors influence the accuracy of the numerical solution. Furthermore, it enables the development of strategies to minimize the approximation error, thus enhancing the overall quality of the numerical method.

By utilizing the moving nodes method to derive this explicit expression, the paper contributes to a deeper understanding of the approximation error in the context of two-point boundary value problems. This understanding is crucial for advancing numerical methods, as it provides a foundation for improving accuracy and reliability in solving complex differential equations. Ultimately, the insights gained from this analysis can inform future research and applications, paving the way for more effective numerical solutions in various scientific and engineering fields.

When a two-point boundary value problem is solved by difference methods, the question of the degree of approximation usually appears. For the closeness of the exact and approximation of the solution, and the quality of the difference scheme are evaluated based on the degree of this parameter. With such an analysis, other parameters (the coefficients of the differential equation) are not explicitly involved in the approximation error

expression. Obtaining an explicit expression for the approximation error makes it possible to analyze it.

Consider the simplest ordinary differential equation with boundary conditions

$$\frac{d^2u}{dx^2} = C, \quad u(0) = 0, \ u(1) = 1 \tag{1}$$

where C is constant.

Create a uniform grid on segments $[0,1]$ with step h. A uniform grid on a segment $x \in [0,1]$ with step h has the form:

$$\bar{\omega}_h = \{x_k = hk, \ k = 0,1,...,N, \ h \cdot N = 1\}$$

Let us replace the second-order derivative by the difference relation [18]:

$$\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} = C,$$
$$1 \le i \le N-1, \ U_0 = 0, \ U_N = 1 \tag{2}$$

Difference scheme (2) traditionally has order $O(h^2)$. However, if we solve system (2) by the Tomas algorithm, we obtain a numerical solution that coincides with the exact analytical solution for any grid steps h at the grid nodes. Those. scheme (2) approximates (1) exactly.

2. Method For Determining Approximation Error

Let we have a differential equation

$$Lu = f, \tag{3}$$

where L is a differential operator, f is a known function, and u is an unknown function. (3) the equation is considered in some domain D with appropriate boundary conditions. The differential equation (3) is replaced by the difference equation [18] :

$$L_h u_h = f_h, \tag{4}$$

where L_h is the difference operator, u_h is the unknown grid function, and f_h is the approximation of the function f at the grid nodes.

Usually, the approximation error is given as [18,19]:

$$Q_h = L_h[u]_h - f_h, \tag{5}$$

where $[u]_h$ is the exact solution of (3) at the grid nodes. Using the Taylor series, from (5) one obtains that, $Q_h = O(h^m)$, where h is the grid step and m is the degree of approximation.

You can determine an explicit approximation error if you use the method of a moving node, which allows you to extend the definition to the entire area D. This allows you to introduce an approximation error like this:

$$R_h = L_h\{u\}_h - f_h. \tag{6}$$

Here $\{u\}_h$ is a predefined continuous function by means of a moveable node. Approximate calculation of the approximation error of type (6) is demonstrated using simple examples.

3. Results and Discussion

As an application of the above approach, consider examples.

3.1. Simple Boundary Value Problem

Consider a simple boundary value problem:

$$\frac{d^2u}{dx^2} = f(x), \quad u(0) = u_a, \ u(1) = u_b \tag{7}$$

Let's build a non-uniform grid on segments $[0,1]$:

$$\bar{\omega}_h = \{0 = x_0 < x_1 < ... < x_{N-1} < x_N = 1, \ k = 0,1,...,N\}$$

In the non-uniform grid, we replace (7) with the difference problem:

$$\frac{2}{x_{i+1} - x_{i-1}}\left(\frac{U_{i+1} - U_i}{x_{i+1} - x_i} - \frac{U_i - U_{i-1}}{x_i - x_{i-1}}\right) = f(x_i), \tag{8}$$
$$i = 1,2,...,N-1.$$

Here U_i is the grid solution of the problem. From here

$$U_i = \frac{U_{i+1}(x_i - x_{i-1}) + U_{i-1}(x_{i+1} - x_i)}{x_{i+1} - x_{i-1}} -$$
$$\frac{1}{2} f(x_i)(x_i - x_{i-1})(x_{i+1} - x_i), \quad i = 1,2,...,N-1. \tag{9}$$

We redefine the value of the function at non-nodal points as follows. To do this, we consider in (9) $x_{i+1}, x_{i-1}, U_{i-1}, U_{i+1}$, to be fixed, and x_i to be moved, and the function $f(x)$ to be smooth. Thus, we will complete the grid function on each segment (x_{i-1}, x_{i+1}). From (9) we get

$$U_i''(x_i) = -\frac{1}{2} f''(x_i)(x_{i+1} - x_i)(x_i - x_{i-1}) - \tag{10}$$
$$f'(x_i)(x_{i+1} + x_{i-1} - 2x_i) + f(x_i)$$

Then the approximation error for the nodal points looks like this:

$$R_h(x_i) = -\frac{1}{2} f''(x_i)(x_{i+1} - x_i)(x_i - x_{i-1}) - \tag{11}$$
$$f'(x_i)(x_{i+1} + x_{i-1} - 2x_i)$$

If the grid is uniform for the approximation error, we obtain the expression

$$R_h(x_i) = -\frac{1}{2} f''(x_i)h^2, \ i = 1, 2, ..., N-1. \quad (12)$$

If on the segments (x_{i-1}, x_{i+1}) the function constant approximation error is identically equal to zero and we get the exact solution.

Based on expression (10), the following conclusion can be drawn.

Given a two-point boundary value problem

$$\frac{d^2u}{dx^2} = f^*(x), \quad u(0) = u_a, \ u(1) = u_b$$

and $f^*(x)$ can be represented as

$$f^*(x_i) = -\frac{1}{2} f''(x_i)(x_{i+1} - x_i)(x_i - x_{i-1}) - f'(x_i)(x_{i+1} + x_{i-1} - 2x_i) + f(x_i)$$

then the difference scheme

$$\frac{2}{x_{i+1} - x_{i-1}} \left(\frac{U_{i+1} - U_i}{x_{i+1} - x_i} - \frac{U_i - U_{i-1}}{x_i - x_{i-1}} \right) = f(x_i), \quad i = 1, 2, ..., N-1,$$

gives a grid solution coinciding with the exact solution at the nodal points.

If there is only one internal node point (the node being moved is one), then an approximate analytical solution can be obtained. Indeed, if we rewrite scheme (8) for one moving node, we have

$$2\left(\frac{U_b - U(x)}{1 - x} - \frac{U(x) - U_a}{x} \right) = f(x_i). \quad (13)$$

From here we obtain an approximate analytical solution:

$$U(x) = U_b x + U_a(1 - x) - \frac{1}{2} f(x_i)(1 - x)x. \quad (14)$$

In this case, (14) represents the exact solution of the problem (7) if we put

$$f^*(x) = -\frac{1}{2} f''(x)(1 - x)x - f'(x)(1 - 2x) + f(x).$$

The form of the approximation error (11) allows the construction of new schemes of the collocation type. Indeed, if in problem (8) we replace the right side by the expression

$$f(x_i) + A(x_i - x_{i-1})(x_{i+1} - x_i),$$

Here A is still an unknown constant. Parameter A is determined so that the approximation error (11) for a uniform step at node x_i is equal to zero, i.e. collocation type scheme. Then we have

$$A = \frac{1}{4} f''(x_i)$$

.

3.2. Boundary value problem for convection and diffusion equation

Consider a stationary equation in which only convection and diffusion are present without a source.

$$\varepsilon u'' + u' = 0, \quad (15)$$

with boundary conditions $v(0) = 0, v(1) = 1$.

There are various schemes for the difference solution (15) [6, 7]. Based on the moving node technique [1,2], it is possible to explicitly express local errors in the approximation of differential equations. Using the moving node method [1], we will show the efficient calculation of local approximation errors for the model problem (15).

3.1.1. Scheme with central-difference approximation of the convective term

Take a segment $[x_{i-1}; x_{i+1}]$ and any point x. Consider the grid analog (15)

$$\frac{2\varepsilon}{x_{i+1} - x_{i-1}} \left(\frac{u_{i+1} - u}{x_{i+1} - x} - \frac{u - u_{i-1}}{x - x_{i-1}} \right) + \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}} = 0 \quad (16)$$

At $x = (x_{i+1} - x_{i-1})/2$, we have a central difference approximation. Here, u_{i+1} is the approximate value of the solution at the point x_{i+1}, u_{i-1} is the approximate value of the solution at the point x_{i-1}.

From (16) we find

$$u = \frac{1}{2\varepsilon(x_{i+1} - x_{i-1})} [(x - x_{i-1})(2\varepsilon + x_{i+1} - x)u_{i+1} + (x_{i+1} - x)(2\varepsilon - x + x_{i-1})u_{i-1}] \quad (17)$$

.

From here we get,

$$u' = \frac{2\varepsilon + x_{i+1} + x_{i-1} - 2x}{2\varepsilon} \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}}, \quad (18)$$

$$u'' = -\frac{1}{\varepsilon} \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}}. \quad (19)$$

If the difference solution at nodal points is known, then formula (17) makes it possible to determine the unknown at points that are not nodal.

Using formulas (18) and (19), the derivatives are restored at any point of the segment. Multiplying (19) by and adding with (18), we obtain

$$\varepsilon u'' + u' = \Psi_1, \quad (20)$$

where

$$\Psi_1 = \frac{x_{i+1} + x_{i-1} - 2x}{2\varepsilon} \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}}.$$

Equation (20) can be called a differential analog of the difference equation (16); difference equation (16) is a collocation-type scheme.

Using (19), the approximation error can be written as

$$\Psi_1 = -\frac{x_{i+1} + x_{i-1} - 2x}{2} u''.$$

Then equation (20) takes the form

$$\left(\varepsilon + \frac{x_{i+1} + x_{i-1} - 2x}{2}\right) u'' + u' = 0. \quad (21)$$

Thus, difference equation (16) exactly approximates differential equation (21) on the segment $[x_{i-1}, x_{i+1}]$.

Comparison of Eqs. (15) and (21) shows that when Eq. (15) is approximated by scheme (16), scheme diffusion appears with a variable coefficient $(x_{i+1} + x_{i-1} - 2x)/2$.

3.2.2 **Upwind Scheme**. Let us consider the difference analog of equation (15), in which the convective term is approximated by the one-sided difference relation

$$\frac{2\varepsilon}{x_{i+1} - x_{i-1}} \left(\frac{u_{i+1} - u}{x_{i+1} - x} - \frac{u - u_{i-1}}{x - x_{i-1}} \right) + \frac{u_{i+1} - u}{x_{i+1} - x} = 0. \quad (22)$$

From here we get

$$u = \frac{(x - x_{i-1})(2\varepsilon + x_{i+1} - x_{i-1})}{(x_{i+1} - x_{i-1})(2\varepsilon + x - x_{i-1})} u_{i+1} + \frac{2\varepsilon(x_{i+1} - x)u_{i-1}}{(x_{i+1} - x_{i-1})(2\varepsilon + x - x_{i-1})} \quad (23)$$

Determine the first and second derivatives:

$$u' = \frac{2\varepsilon(2\varepsilon + x_{i+1} - x_{i-1})}{(2\varepsilon + x - x_{i-1})^2} \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}}, \quad (24)$$

$$u'' = \frac{-4\varepsilon(2\varepsilon + x_{i+1} - x_{i-1})}{(2\varepsilon + x - x_{i-1})^3} \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}} \quad (25)$$

Let us calculate the approximation error

$$\Psi_2 = \frac{2\varepsilon(x - x_{i-1})(2\varepsilon + x_{i+1} - x_{i-1})}{(2\varepsilon + x - x_{i-1})^3} \frac{u_{i+1} - u_{i-1}}{x_{i+1} - x_{i-1}}$$

The differential analog of scheme (22) has the form

$$\left(\varepsilon + \frac{x - x_{i-1}}{2}\right) u'' + u' = 0, \quad (26)$$

those. with a scheme against the flow, we have a scheme diffusion with a coefficient . Based on (23) - is a hyperbola, which is monotone on the segment, i.e. scheme (22) is monotonic.

Based on the form of the differential analogue (26), we can conclude that the differential equation

$$\left(\varepsilon + \frac{x}{2}\right) u'' + u' = 0 \quad (27)$$

is exactly approximated by the scheme

$$2\varepsilon \left(\frac{u_b - u}{1 - x} + \frac{u - u_a}{x} \right) + \frac{u_b - u}{1 - x} = 0 \quad (28)$$

Those. solving (28) with respect to u, we obtain the exact solution of differential equation (27).

3.3. Parametric Schemes

In this case, an attempt is made to create a special parametric scheme in order to improve the quality of the circuit. The peculiarity of this approach is the choice of the parameter, which is carried out on the basis of the calculated approximation error, which allows more accurately adjusting the parameters of the scheme to achieve the best indicators. We demonstrate the effectiveness of this method using examples of problems related to convection-diffusion processes, where the correct choice of parameters is especially important for the stability and accuracy of the solution. Consider the problem [19,20].

$$Pe\frac{du}{dx} = \frac{d^2u}{dx^2} + Pe \cdot S(x), \quad u(0) = u_0, \quad u(1) = u_1, \quad (29)$$

Here Pe is the Peclet number, $S(x)$ is the source, u is the unknown function.

When problem (29) is discredited, it is essential to approximate the convective term [4]. The standard finite-difference scheme against the flow on a three-point template is:

$$Pe\frac{U - U_W}{x - x_W} = \frac{2}{x_E - x_W} \left(\frac{U_E - U}{x_E - x} - \frac{U - U_W}{x - x_W} \right) + Pe \cdot S(x), \quad (30)$$

Consider the parametric scheme

$$Pe\frac{U - U_W}{x^k - x_w^k} \cdot kx^{k-1} = \frac{2}{x_E - x_W} \left(\frac{U_E - U}{x_E - x} - \frac{U - U_W}{x - x_W} \right) + Pe \cdot S(x), \quad (31)$$

The choice of the parameter k can be found by numerical experiment. Based on the calculated approximation error R_h, it is not difficult to select the parameter k. The idea of approximating the convective term is as follows. We introduce an intermediate variable $y(x)$, and based on the calculation of the derivative of a complex function, we have

$$\frac{du}{dx} = \frac{du}{dy} \cdot \frac{dy}{dx}.$$

For the function $y(x)$ we take a monotonically increasing function, for example, $y = x^k$. du/dy will

be replaced by the difference relation upstream. Making the assumption that with such a replacement, the approximation error decreases. In this way

$$\frac{du}{dx} \approx \frac{u-u_W}{x^k - x_w^k} \cdot kx^{k-1}.$$

Figure 1 shows the results of calculations $(Pe=0,\ S(x)=0,\ N=11, u_0=0,\ u_1=1)$, at $k=1$ and $k=9$.

Thus, by carefully choosing the parameter k, we are able to obtain a result that is as close as possible to the exact solution of the problem. This approach allows us to significantly increase the accuracy and reliability of calculations, minimizing approximation errors and ensuring more stable behavior of the numerical method.

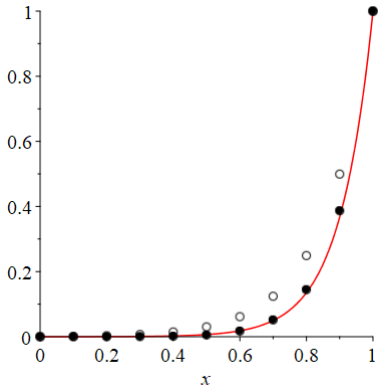

Figure 1: Comparison of results. The solid line is the exact solution, the circles are the numerical results obtained at k=1, and the solid circles at k=9.

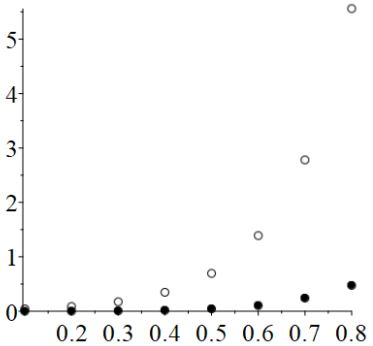

Figure 2: Comparison of the results of the approximation error at internal nodal points. The solid circles are obtained according to the scheme (31) at k=9, and the circles at k=1.

3.3. Iterative method to get a solution

It is known that after replacing the differential equation with discrete ones, we obtain a system of algebraic equations [4,5,19,20]. There are two approaches to solving systems of algebraic equations: exact methods and iterative methods. Using the idea of constructing iterative methods for systems of discrete equations, we will show the possibilities of an analytical approximate solution based on the method of moving nodes.

Consider problem (29). If there is only one moving node, approximating the convective term by the upstream scheme from (31) we get $(u_0=0,\ u_1=1)$.

$$u^1 = \frac{2x}{2+Pe(1-x)}. + \frac{x(1-x)}{2+Pe(1-x)} \cdot S(x) \qquad (32)$$

This expression is taken as the initial approximation of problem (29). Let's find the approximation error

$$R^1 = \frac{d^2u^1}{dx^2} - Pe\frac{du^1}{dx} + Pe \cdot S(x) \qquad (33)$$

Let's calculate the second approximation

$$u^2 = u^1 + \omega x(1-x)R^1$$

Find the approximation error R^2.

$$R^2 = \frac{d^2u^2}{dx^2} - Pe\frac{du^2}{dx} + Pe \cdot S(x)$$

Thus, we carry out an iterative process in the form

$$u^k = u^{k-1} + \omega x(1-x)R^{k-1} + Pe \cdot S(x),\ \ k=2,3\ldots \qquad (34)$$

In (34) ω is the relaxation parameter.

In Fig. 3 the exact solution of the problem as well as approximating analytical solutions u^1, u^2, u^3 and u^4 are compared. As can be seen from the graphic, step by step we can improve of analytical solution $(S(x)=0,\ Pe=10,\ \omega=0.08)$.

On fig. 4 the sequence of solution of problem (18) is given for $S(x)=\cos(5x),\ Pe=10,\ \omega=0.06$. On fig. 3 and 4, the solid line corresponds to the exact solution of the problem; dot - u^1; dashed, u^2; ; dotted-dashed -- u^3; long-dashed - u^4.

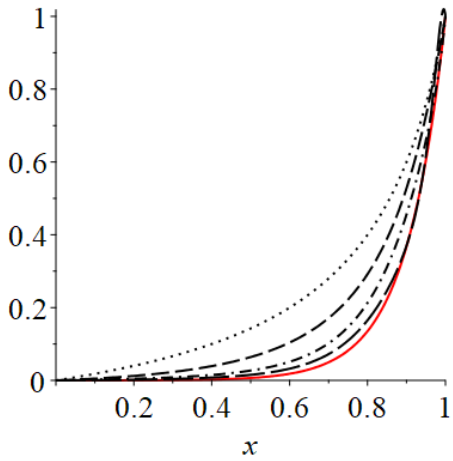

Figure 3: Comparison of results: S(x) = 0, Pe=10, ω=0,08

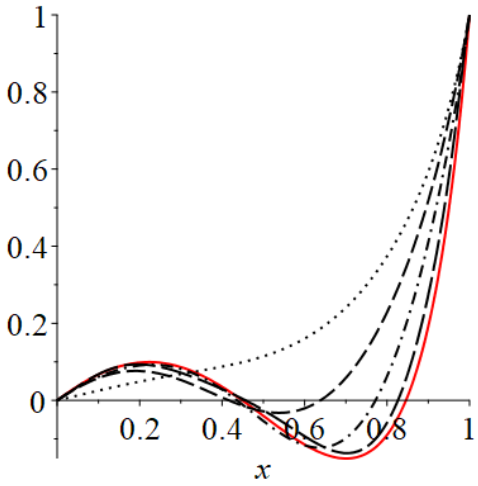

Figure 4: Comparison of results: S(x) = cos(5x), Pe=10, ω=0,06

As can be seen from the graphic, step by step we can improve of the analytical solution.

References

[1] U. Dalabaev and D. Khasanova, "An explicit expression of ordinary difference schemes for differential equations by the moved node method," AIP Conf. Proc., vol. 3004, no. 1, p. 060043, 2024. doi: 10.1063/5.0145881.

[2] G. D. Smith, Numerical Solution of Partial Differential Equations: Finite Difference Methods, 3rd ed. Oxford, UK: Oxford University Press, 1985. ISBN: 978-0198596509.

[3] R. D. Richtmyer and K. W. Morton, Difference Methods for Initial-Value Problems, 2nd ed. New York, NY, USA: Wiley-Interscience, 1967. ISBN: 978-0470720400.

[4] S. V. Patankar, Numerical Heat Transfer and Fluid Flow. Washington, DC, USA: Hemisphere Publishing Corporation, 1980. ISBN: 978-0070487406.

[5] A. A. Samarskii, Introduction to the Theory of Difference Schemes. Moscow, Russia: Nauka, 1971. (In Russian).

[6] R. E. Mickens, Nonstandard Finite Difference Models of Differential Equations. Singapore: World Scientific, 1994. ISBN: 978-9810214586.

[7] R. E. Mickens, "Calculation of denominator functions for nonstandard finite difference schemes for differential equations satisfying a positivity condition," Numer. Methods Partial Differ. Equ., vol. 23, no. 3, pp. 672–691, 2007. doi: 10.1002/num.20198.

[8] R. E. Mickens, "Exact solutions to a finite-difference model of a nonlinear reaction-advection equation: Implications for numerical analysis," J. Differ. Equ. Appl., vol. 8, no. 9, pp. 823–847, 2002. doi: 10.1080/1023619021000037086.

[9] E. M. Adamu, K. C. Patidar, and R. R. Mickens, "An unconditionally stable nonstandard finite difference method to solve a mathematical model describing visceral leishmaniasis," Math. Comput. Simul., vol. 187, pp. 171–190, 2021. doi: 10.1016/j.matcom.2021.03.006.

[10] M. E. S. Begaray-Fesquet and B. B. Garay-Fesquet, "Extending nonstandard finite difference schemes rules to systems of nonlinear ODEs with constant coefficients," Math. Numer. Anal., vol. 12, 2021.

[11] A. A. Ç. Köroğlu, "Exact and nonstandard finite difference schemes for the Burgers equation B(2,2)," Turk. J. Math., vol. 45, pp. 647–660, 2021.

[12] D. U. Dalabaev, "Difference analytical method of the one-dimensional convection-diffusion equation," Int. J. Innov. Sci. Eng. Technol., vol. 3, pp. 234–239, 2016.

[13] D. U. Dalabaev, "Computing technology of a method of control volume for obtaining of the approximate analytical solution to one-dimensional convection-diffusion problems," Open Access Library J., vol. 5, p. e504962, 2018.

[14] U. Dalabaev and R. Abdurakhmanov, "A simple way to solve boundary value problems in technological processes," J. Appl. Math. Comput., vol. 23, no. 4, pp. 456–462, 2023.

[15] D. U. Dalabaev and X. D. X., "The approximation error of ordinary differential equations based on the moved node method," Probl. Comput. Appl. Math., vol. 5, no. 24, pp. 5–9, 2022.

[16] D. U. Dalabaev, "Application of the method of moving nodes to solving applied boundary value problems," Bull. Inst. Math., vol. 6, pp. 5–9, 2018.

[17] R. Abdurakhmanov and D. U. Dalabaev, "Computational technology for improving the quality of difference schemes based on moving nodes," J. Comput. Math. Appl., vol. 1860, pp. 112–118, 2021.

[18] S. A. V. P. N., Numerical Methods for Resolving Convection-Diffusion Problems. Moscow, Russia: Book House "LBROKOM", 2015.

[19] S. A. A. V. B., Difference Methods for Elliptic Equations. Moscow, Russia: Nauka, 1976.

[20] S. A. N. E. S., Methods for Solving Grid Equations. Moscow, Russia: Nauka, 1978.

DALABAEV UMURDIN has done his bachelor's degree from National University of Uzbekistan in 1969. He has done his PhD degree from Institute of Mechanics of the Academy of Sciences of Uzbekistan in 1976. He has completed his DSc degree from National University of Uzbekistan in 2021.

KHASANOVA DILFUZA has done her bachelor's degree from Andijan State University in 2004. She has done her master's degree from Andijan State University in 2006.

.

JENRS

Received: 17 May, 2025, Revised: 14 June, 2025, Accepted: 13 June, 2025, Online: 26 June, 2025

DOI: https://doi.org/10.55708/js0406003

Comparative Analysis of Supervised Machine Learning Models for PCOS Prediction Using Clinical Data

Ranyah Taha*[,1], Huda Zain El Abdin [2], Tala Musleh [3]

[1] Computer Science Dept., Al-Iman School, Bahrain

[2] Faculty of Science and Technology, Computer Science Department, University of Middlesex, London, Hendon, United Kingdom

[3] Pharmacy Department, College of Health and Sports Sciences, University of Bahrain, Bahrain

Corresponding author: Huda Zain El Abdin, Wokingham, United Kingdom, hudaa.z@icloud.com , +447832032740

ABSTRACT: Polycystic Ovary Syndrome (PCOS) is a prevalent hormonal disorder affecting women of reproductive age, commonly resulting in irregular menstrual cycles, elevated androgen levels, and the presence of polycystic ovaries. It is a major cause of infertility and is often linked with metabolic complications such as insulin resistance and obesity. Symptoms vary and may include acne, excessive hair growth, weight gain, and hair thinning. Early detection and proper management through lifestyle interventions and medical treatment are crucial to mitigating long-term health risks. This study investigates the classification performance of seven supervised machine learning algorithms—Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Classifier (GBC), Adaptive Boosting (AdaBoost), and Multi-Layer Perceptron (MLP)—using clinical and lifestyle data related to PCOS. The models were evaluated using accuracy, precision, recall, F1 score, and ROC AUC metrics. LR consistently outperformed the other models, achieving the highest accuracy (91.7%), precision (96%), and Receiver Operating Characteristics -Area Under the Curve (ROC AUC) (96.8%), while also maintaining a strong balance in recall and F1 score. This outstanding performance is attributed to the linear nature of the dataset and the efficiency, simplicity, and generalizability of LR, making it particularly suitable for this classification task. This study introduces a novel approach for predicting PCOS by integrating advanced data preprocessing techniques with a focus on model simplicity and interpretability. The predictive performance of LR was further enhanced through the application of the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance and Analysis of Variance (ANOVA) F-score-based feature selection to identify the most statistically significant predictors. This approach not only achieved high predictive accuracy but also ensured transparency and ease of deployment, making it highly applicable for clinical decision-support systems aimed at early and accurate PCOS diagnosis.

KEYWORDS: Artificial Intelligence, Data Analysis, Polycystic Ovary Syndrome, Supervised Machine Learning, Medical Diagnosis.

1. Introduction

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder that affects approximately 8–13% of women of reproductive age worldwide. Its prevalence, however, varies depending on ethnicity and diagnostic criteria. PCOS is characterized by hormonal imbalances, particularly elevated androgen levels, which manifest in symptoms such as irregular menstrual cycles, anovulation, and the presence of multiple ovarian follicles. These disruptions often lead to infertility and are commonly accompanied by metabolic complications, including obesity, insulin resistance, type 2 diabetes, and increased cardiovascular risk [1]. Despite its widespread occurrence and clinical implications, PCOS remains underdiagnosed due to its heterogeneous presentation and overlapping symptoms with other conditions. This diagnostic challenge underscores the need for advanced tools to enhance early detection and personalized care. Artificial intelligence (AI), particularly machine learning (ML), offers a promising solution by identifying complex, non-linear patterns within clinical and biochemical data—patterns that may be overlooked through conventional diagnostic approaches. Numerous studies have highlighted the potential of ML to augment clinical

workflows in endocrinology, providing timely, data-driven support for healthcare professionals [1].

The presence of PCOS symptoms can vary significantly among women, it includes acne, extra body hair growth, hair thinning and obesity. As a matter of fact, symptoms differ from one person to another, which makes diagnosing PCOS challenging. Early and accurate diagnosis is essential for timely intervention to manage both reproductive and metabolic health risks [2].

The diagnostic process for PCOS remains challenging due to the diverse presence of symptoms across women. Typically, physicians rely on a combination of clinical assessments, blood tests, and pelvic ultrasound imaging. However, the absence of a comprehensive diagnostic tool also makes it hard to distinguish from other conditions leading to misdiagnosis or delays in diagnosis. Consequently, healthcare systems are seeking more advanced solutions to boost diagnostic accuracy to have efficient outcomes [3].

Advancements in AI, specifically in ML, have shown considerable promise in healthcare diagnostics. This is particularly for finding and classifying sophisticated diseases such as PCOS. For instance, supervised learning algorithms are demonstrating significant capability by uncovering hidden patterns within clinical and lifestyle data that was not readable by healthcare providers. The availability of electronic health records (EHRs) and patient data are rapidly increasing. Furthermore, AI-driven solutions could improve the prediction of PCOS diagnosis, enabling tailored and patient-specific treatments [4].

This study investigates the effectiveness of seven ML algorithms—LR, NB, SVM, RF, GBC, AdaBoost and MLP—in identifying PCOS using a dataset sourced from Kaggle. Following the CRISP-DM framework, the study applies a structured approach to data analysis and model development, incorporating patient data related to symptoms and lifestyle factors. The performance of each model is assessed using precision, recall, F1 score, and ROC AUC to enable a comparative evaluation of their strengths and limitations.

The findings aim to inform the development of AI-based diagnostic tools that support clinicians in diagnosing PCOS more accurately and efficiently, thereby enhancing clinical decision-making.

The study is structured into the following sections: literature review, methodology, data description and preprocessing, model implementation, results, discussion, conclusion, and future recommendations.

2. Literature Review

Several studies in recent years have used different ML techniques to diagnose and predict PCOS. Utilising clinical and physiological dataset to augment prediction accuracy. These approaches enhance distinct algorithms and data preprocessing methods for the aim of capturing patterns that assist in early and reliable PCOS detection.

In [5], the Decision Tree (DT), RF, and SVM algorithms were applied to a clinical dataset containing features such as Body Mass Index (BMI), insulin levels, and follicle count to predict the presence of PCOS. Among the models tested, the RF classifier achieved the highest accuracy of 89.5%. The study emphasized that ensemble models like RF are particularly effective in capturing complex relationships and interdependencies among clinical features.

Similarly, authors [6] used LR, NB, and KNN to analyse a dataset of 520 PCOS cases. In terms of model performance development, the study focused on feature selection techniques such as chi-square and recursive feature elimination. LR revealed strong predictive capability with an accuracy of 85.3%, especially when hormonal and metabolic attributes were emphasized. This demonstrates the strength of tree-based models in the clinical field.

In a more recent analysis, authors in [7] implemented DL models accompanied with traditional supervised classifiers on a refined clinical dataset. The study compared Artificial Neural Networks (ANN) with SVM, DT, and XGBoost. Despite the fact that ANN achieved the highest accuracy of 91.2%, the authors highlighted that simpler supervised model like XGBoost provided competitive results with lower computational costs, supporting their practicality for clinical integration.

In the imaging domain, researchers [8] proposed a model interpretability by combining DT classifiers with SHapley exPlanations (SHAP), a method that collaborates each independent feature to contribute to accurate predictions. This approach assembled the authors to generate a ranked list of features based on their impacts on the model's output. Nevertheless, testosterone levels and the luteinizing hormones (LH) to follicle-stimulating hormone (FSH) ratio emerged as dominant predictors lining up with clinical indicators of PCOS. Through the visualization of feature importance at both the population and patient-specific levels, the study provided a clearer understanding of the model's reasoning, which contributes to greater clinical confidence and interpretability in automated diagnostic applications.

Furthermore, authors [9] established a cloud-based diagnostic system trained on three different medical datasets taken from medical centres. AI algorithms analysed images, focusing on DNA content with cell nuclei. It validated the value of feature specificity such as DNA content as PCOS markers. Based on the results, these images were derived to a cloud-based platform for

evaluation and assessments. Results achieved accuracy between 86% and 89%.

In addition, Arya [10] proposed a two-step approach to medical diagnosis that merges both supervised and unsupervised learning techniques. Starting with k-means clustering was used to group similar patient records. Followed by, analysing the clustered groups using supervised classification models, DT and SVM models to predict diagnosis. This combined method improved the accuracy of the system, reaching a prediction accuracy of 87.5%, and highlighted how blending ML techniques.

In the use of Graph Neural Network (GNN), Boll, et al. [11] acknowledges relationships between variables in EHRs. By treating clinical data as a network each variable is a node, and the connections reflect how these variables interact. As a result, patient information was modelled in a meaningful way. This graph-based approach achieved a strong AUC score of 89%, showing significant clinical prediction outcomes using advanced Deep Learning (DL) techniques.

Similarly, authors in [12] developed a Light Gradient Boosting Machine (LightGBM) model in conjunction with SHAP to identify and prioritise features relevant to PCOS diagnosis. The analysis highlighted the significance of anti-Müllerian hormone (AMH) levels and clinical signs such as hirsutism in prediction PCOS. As a result, the model achieved AUC of 93%, indicating high performance. In another notable comparison, Wang, et al. [13] implemented SVM, GBC, and MLP on PCOS datasets with categorical and numerical features. MLP achieved the highest F1 score 92%, demonstrating DL's ability to capture nonlinear relationships in diverse data formats. However, SVM maintained excellent generalization with less overfitting.

Additionally, authors in [14] examined the performance of LR, SVM, and MLP for early PCOS detection using lifestyle data (e.g., activity, sleep). Results show LR proved superior in AUC and interpretability, confirming its dominance in structured health data settings. Specifically, the study documented an AUC of 82.3% for the LR model, highlighting its robust performance.

Addressing the challenge of class imbalance, authors in [15] conducted an analysis on distinct algorithms, RF AdaBoost, and GBC on datasets with imbalanced PCOS class distributions. By applying SMOTE for balance, GBC performed best in handling rare class detection, with an AUC of 94.2%, followed closely by AdaBoost.

Similarly, authors in this study [16] developed predictive models using four ML methods: LR, SVM, GBC trees, and RF. It focused on hormone values (follicle-stimulating hormone, luteinizing hormone, oestradiol, and sex hormone-binding globulin) were combined to create a multilayer perceptron score using a neural network classifier. The models achieved AUC values of 85%, 81%, 80%, and 82%, respectively. Significant positive predictors of PCOS diagnosis across models included hormone levels and obesity; negative predictors included gravidity. The study illustrates the potential benefits of integrating AI tools into EHRs to facilitate earlier detection of PCOS.

Finally, researchers in [17] proposed three lightweight DL models LSTM-based, CNN-based, and CNN-LSTM-based for automated PCOS prediction. To address the imbalanced nature of the dataset, the SMOTE was employed. The models achieved accuracies of 92.04%, 96.59%, and 94.31%, with corresponding ROC-AUC values of 92.0%, 96.6%, and 94.3%. The study highlights the effectiveness of lightweight DL models in delivering high performance with fewer trainable parameters, making them suitable for resource-constrained environments.

Previous studies have utilized various ML algorithms to enhance PCOS diagnosis and prediction. Among these, RF demonstrated strong predictive capabilities by capturing complex, non-linear relationships, achieving accuracies up to 89.5%. LR was also widely used due to its simplicity and interpretability, particularly effective with structured clinical and lifestyle data, achieving accuracies above 85%.

SVM provided good generalization performance, especially on smaller datasets, but was sometimes outperformed by DL models on larger datasets. DL approaches, including ANN, CNN, and LSTM, achieved the highest accuracies, reaching up to 96.59% with CNN-LSTM architectures, though they required higher computational resources.

Tree-based ensemble models such as GBC and XGBoost delivered competitive results with lower computational costs, making them suitable for clinical environments. GBC particularly excelled in handling imbalanced datasets, achieving AUC values over 94%. Recently, advanced models like GNN were introduced to model complex relationships in electronic health records, achieving an AUC of 89%.

In summary, although DL models achieved the highest prediction accuracies, RF and GBC provided a balanced trade-off between performance, interpretability, and computational efficiency, making them highly applicable in practical clinical scenarios.

3. Research Methodology and approach

3.1. Background of the Research Study

This research was conducted using Google Colab as the primary development environment, with Scikit-learn

as the main Python library for implementing ML models. A total of seven classification algorithms were employed to analyse and classify PCOS cases. The models used include LR, NB, SVM, RF, GB, AdaBoost, and MLP. Each algorithm was trained and evaluated to assess its effectiveness in accurately identifying PCOS based on clinical and lifestyle features.

The selection of these specific algorithms—LR, NB, SVM, RF, GBC, AdaBoost, and MLP—was driven by their complementary strengths in handling structured clinical data. LR offers high interpretability and computational efficiency, making it ideal for linear relationships within medical datasets. NB is well-suited for smaller datasets and performs effectively under the assumption of conditional feature independence. SVM is robust in high-dimensional spaces and generalizes well across complex boundaries. Ensemble methods such as RF, GBC, and AdaBoost are powerful in modeling non-linear interactions and addressing class imbalance, which are common in PCOS-related data. Lastly, MLP, a type of artificial neural network, was included for its ability to capture deep non-linear relationships. This diverse algorithm selection enables a comprehensive comparison across linear, probabilistic, ensemble-based, and neural learning paradigms, enhancing the model's applicability to the multifactorial nature of PCOS.

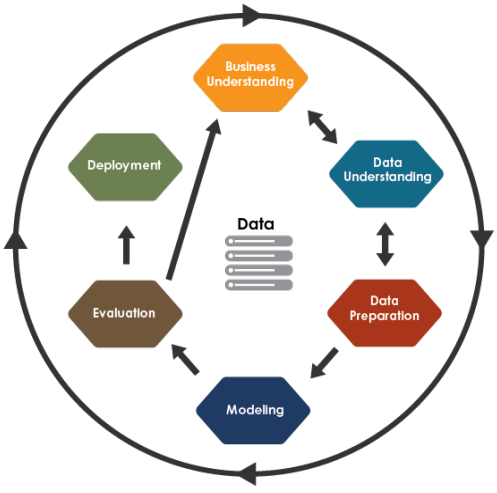

Figure 1: Phases of the CRISP-DM Methodology

The study followed the CRISP-DM methodology, a widely recognized framework for structuring ML projects. This approach consists of six key phases: defining project objectives (business understanding), exploring and analysing the dataset (data understanding), organizing and cleaning data for analysis (data preparation), developing and tuning ML models (modelling), assessing model performance (evaluation), and preparing the model for practical application (deployment) [18]. Adopting this structured workflow ensured clarity, consistency, and effectiveness throughout the project, ultimately contributing to the reliable and accurate results presented in Figure 1.

3.2. Dataset Description

The dataset used in this study was retrieved from Kaggle, a widely recognized platform for data science competitions and open-access datasets [19]. It contains clinical, biochemical, and lifestyle-related information collected from 541 female patients to support the prediction and diagnosis of PCOS. The dataset includes 44 features, including a binary target variable, PCOS (Y/N), where a value of 1 indicates a confirmed diagnosis of PCOS and 0 denotes its absence.

The features span several categories. Demographic and anthropometric variables include age, weight, height, BMI, and blood group. Vital signs such as pulse rate, respiratory rate, and blood pressure are included. Reproductive health indicators—like menstrual cycle regularity and pregnancy status—are complemented by hormonal measurements including AMH, FSH, LH, the FSH/LH ratio, and Beta-HCG. The dataset also captures symptoms and lifestyle factors, such as hair loss, acne, skin pigmentation, weight gain, hirsutism, fast food intake, and physical activity. Furthermore, ultrasound features detail follicle count and size in each ovary, along with endometrial thickness.

Notably, this dataset does not contain some of the core hormonal biomarkers typically used in the clinical diagnosis of PCOS, such as estrogen, progesterone, and testosterone. The absence of these indicators constitutes a key limitation of the dataset provided via Kaggle and was not a modeling decision but rather a constraint imposed by data availability. In real-world clinical practice, these hormones are fundamental to differential diagnosis and are often among the first parameters assessed alongside imaging. Their exclusion may restrict the model's ability to fully replicate the diagnostic reasoning employed by clinicians and can limit generalizability to broader patient populations. Future studies will aim to incorporate such hormonal data to enhance both predictive performance and clinical validity.

Additionally, the dataset does not include crucial demographic attributes such as ethnicity, geographical origin, and socioeconomic status—factors that significantly influence hormonal expression, symptomatology, and PCOS risk profiles. The lack of these variables introduces potential bias and restricts the fairness and applicability of the model across diverse populations. This limitation will be acknowledged explicitly in the revised manuscript, and future research will seek to mitigate these shortcomings through more inclusive and representative datasets. A summary of the dataset's attributes is provided in Table 1.

Table 1: Dataset Description

Feature	Description	Data Type
Age (yrs)	Age of the patient in years	Float64
Weight (Kg)	Body weight in kilograms	Float64
Height (Cm)	Height in centimetres	Float64
BMI	Body Mass Index	Float64
Blood Group	Blood type as numerical code	Int64
Pulse rate(bpm)	Pulse rate in beats per minute	Float64
RR (breaths/min)	Respiratory rate per minute	Int64
Cycle(R/I)	Menstrual cycle regularity	Int64
Pregnant(Y/N)	Pregnancy status (1=Yes, 0=No)	Int64
I beta-HCG (mIU/mL)	Beta-HCG hormone level (case I)	Float64
AMH (ng/mL)	Anti-MÃ¼llerian Hormone level	Float64
FSH (mIU/mL)	Follicle Stimulating Hormone	Float64
LH (mIU/mL)	Luteinizing Hormone	Float64
FSH/LH	Ratio of FSH to LH	Float64
Hair loss(Y/N)	Presence of hair loss (1=Yes, 0=No)	Int64
Skin darkening (Y/N)	Presence of skin pigmentation (1=Yes, 0=No)	Int64
Weight gain(Y/N)	Reported weight gain (1=Yes, 0=No)	Int64
Hair growth(Y/N)	Excessive hair growth (1=Yes, 0=No)	Int64
Pimples(Y/N)	Presence of pimples/acne (1=Yes, 0=No)	Int64
Fast food (Y/N)	Fast food consumption (1=Yes, 0=No)	Float64
Reg.Exercise(Y/N)	Engagement in regular exercise (1=Yes, 0=No)	Int64
Follicle No. (L)	Number of follicles in left ovary	Int64
Follicle No. (R)	Number of follicles in right ovary	Int64
Avg. F size (L) (mm)	Average follicle size in left ovary	Float64
Avg. F size (R) (mm)	Average follicle size in right ovary	Float64
Endometrium (mm)	Thickness of the endometrial lining	Float64
PCOS (Y/N)	Diagnosis of PCOS (1=Yes, 0=No)	Int64

3.3. Dataset Preparation

After completing the data exploration phase, the dataset undergoes a comprehensive preprocessing stage. This phase includes handling missing values, eliminating duplicate records, applying normalization, selecting relevant features, encoding categorical variables, and splitting the data into training and testing sets. These preprocessing steps are crucial to ensure the dataset is clean, well-structured, and suitable for accurate modelling and further analysis.

3.3.1. Missing Data

To ensure the integrity of the dataset, two standard validation functions were applied: isnull (). sum () and duplicated (). sum (). For instance, the isnull (). sum () function was used to detect and count missing values across all columns, while duplicated().sum() identified any repeated rows that could affect data quality. The results confirmed that the dataset contained no missing values or duplicate entries, indicating a high level of completeness and consistency. This verification step is essential, as clean and reliable data forms the foundation for developing accurate and robust ML models.

3.3.2. Balancing the Dataset

The dataset comprises a total of 541 patient records, each containing clinical, biochemical, and lifestyle-related information relevant to the diagnosis of PCOS. The target variable, PCOS (Y/N), is binary, where 1 indicates a positive PCOS diagnosis and 0 indicates the absence of the condition as presented in Figure 2. To address this imbalance and improve the performance of ML models, the study employed SMOTE. The SMOTE generates synthetic examples of the minority class (PCOS) to create a more balanced dataset. This technique helps reduce bias toward the majority class during model training, leading to more reliable and generalizable classification outcomes [17].

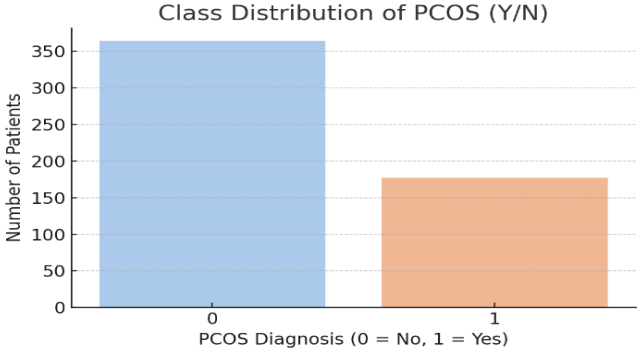

Figure 2: Class Distribution of PCOS

3.3.3. Feature Selection

The feature selection results using ANOVA F-scores highlight the most statistically significant variables for distinguishing between PCOS and non-PCOS cases. The two most predictive features are Follicle No. (R) and Follicle No. (L), with F-scores of 390.84 and 308.52, respectively. These findings are consistent with clinical criteria, as women with PCOS typically present with a higher number of ovarian follicles, particularly in the right ovary. Other highly discriminative features include skin darkening, hair growth, and weight gain, all of which are common symptoms associated with hormonal imbalance and insulin resistance in PCOS patients.

The menstrual cycle regularity feature (Cycle R/I) also shows a high F-score (103.67), emphasizing its importance, as irregular cycles are a key diagnostic marker of PCOS. Moderate contributions come from features like fast food consumption, pimples, weight, BMI, and cycle length, which reflect both lifestyle and physiological factors influencing the condition. Less predictive but still relevant features include hair loss, age, waist size, and hip circumference, which contribute to the model with lower F-scores. Overall, the analysis confirms that reproductive indicators, clinical symptoms, and lifestyle behaviours play a vital role in the classification of PCOS, guiding both feature prioritization and model development for improved diagnostic accuracy. A summary of the attribute's importance is provided in Table 2.

Table 2: Feature Importance Using ANOVA F-score

Selected Feature	ANOVA F-score
Follicle No. (R)	390.83
Follicle No. (L)	308.51
Skin darkening (Y/N)	157.67
hair growth(Y/N)	148.42
Weight gain(Y/N)	130.16
Cycle(R/I)	103.67
Fast food (Y/N)	89.72
Pimples(Y/N)	48.04
Weight (Kg)	25.34
BMI	22.34
Cycle length(days)	17.73
Hair loss(Y/N)	16.6
Age (yrs)	15.75
Waist(inch)	15
Hip(inch)	14.58

3.3.4. Encoding Categorical Data

The dataset was processed using label encoding to convert categorical variables into numerical format, a crucial preprocessing step as most ML algorithms requires numerical input [20]. In this study, all categorical features were successfully transformed into numeric values. This conversion was essential to ensure compatibility with the classification models, ultimately enhancing the efficiency and accuracy of the training and evaluation processes.

3.3.5. Splitting Data

The dataset was initially divided into two subsets, with 80% allocated for training and 20% for testing. This split enables the model to learn patterns from the larger portion of the data while using the remaining portion to assess its performance on previously unseen instances, ensuring a more reliable evaluation.

3.3.6. Data Normalization

The numerical features were normalized to scale their values within a consistent range, typically between 0 and 1. This process ensures that all features contribute equally during model training, preventing any single variable from dominating the learning process. Normalization supports more balanced and unbiased model performance, ultimately enhancing the accuracy and stability of the results [21].

3.4. Modelling

Seven ML algorithms—LR, NB, SVM, RF, GBC, AdaBoost, and MLP—were applied to classify patients based on the presence or absence of PCOS.

LR is a supervised ML algorithm commonly used for binary classification tasks. It estimates the probability that a given input belongs to a particular class by applying a sigmoid function to a linear combination of the input features. The output is a value between 0 and 1, representing the likelihood of the positive class. LR is valued for its simplicity, interpretability, and efficiency, making it a reliable choice for solving classification problems in various domains [22].

RF is an ensemble ML method that constructs numerous DTs during the training phase and combines their predictions to enhance accuracy and stability. For classification tasks, it typically uses majority voting to determine the final output. This approach helps reduce both overfitting and variance compared to relying on a single DT, leading to improved generalization and performance on new, unseen data [23].

GBC is an effective ensemble learning method that constructs models in a sequential manner, with each new model aiming to improve upon the errors of its predecessors. It combines multiple weak learners, typically shallow DTs, and optimizes performance by minimizing a loss function through gradient-based techniques. This approach often results in high predictive accuracy, although it may require more training time due to its iterative nature [23].

SVM is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that separates data points of different classes with the maximum margin. The data points closest to the hyperplane, known as support vectors, are critical in defining the decision boundary. SVM is especially effective in high-dimensional spaces and can be adapted to non-linear problems through the use of kernel functions. Its ability to handle complex relationships and avoid overfitting makes it a widely used method in ML [20].

NB is a simple, yet effective supervised classification algorithm based on Bayes' Theorem. It assumes that all features are independent of each other given the class label—an assumption known as "naive" independence. Despite this simplification, NB performs well in many real-world scenarios, particularly with large datasets. It is computationally efficient, easy to implement, and works well for both binary and multi-class classification problems, especially when the input features are categorical or conditionally independent [20].

AdaBoost is an ensemble learning algorithm that combines multiple weak classifiers, typically DTs, to form a strong classifier. It works by training models sequentially, where each new model focuses more on the errors made by the previous ones. During the training process, weights are assigned to each instance, increasing for those that are misclassified, so the next model gives them more attention. AdaBoost is known for improving accuracy, reducing bias, and being relatively resistant to overfitting when properly tuned. It performs well on binary classification tasks and is particularly effective with clean, well-prepared data [24].

MLP is a type of ANN used for supervised learning tasks, including both classification and regression. It consists of an input layer, one or more hidden layers, and an output layer, with each layer made up of interconnected nodes (neurons).

MLP uses non-linear activation functions and is trained using backpropagation to minimize prediction errors. It is capable of capturing complex patterns in the data but often requires careful tuning of hyperparameters and sufficient data to perform effectively. MLP is particularly useful when the relationship between features and outcomes is non-linear and not easily captured by simpler models [25].

3.5. *Performance Evaluation*

The performance of the supervised ML models is assessed using key evaluation metrics—accuracy, precision, recall, F-measure and ROC AUC—which together offer a comprehensive understanding of each model's classification effectiveness.

3.5.1. *Accuracy:*

It measures the proportion of correctly predicted instances out of the total number of predictions. It reflects the overall effectiveness of a model incorrectly classifying both positive and negative cases, as expressed in Equation (1) [26].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

3.5.2. *F-measure:*

It provides a balanced evaluation by combining precision and recall into a single metric. It is especially valuable when dealing with imbalanced datasets or when both false positives and false negatives carry significant consequences, as shown in Equation (2) [26].

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (2)$$

3.5.3. *Precision:*

It quantifies the ratio of correctly predicted positive instances to all instances predicted as positive. It evaluates the model's ability to produce reliable positive predictions, helping determine how many of the predicted positives are relevant. This is illustrated in Equation (3) [26].

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

3.5.4. *Recall:*

It measures the proportion of actual positive cases that are correctly identified by the model. It is crucial in contexts where missing positive cases may have serious implications, as represented in Equation (4) [26].

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

3.5.5. *ROC AUC*

It is a performance metric used to evaluate the classification ability of a ML model across various threshold settings. The ROC curve plots the True Positive Rate against the False Positive Rate, showing how the model's sensitivity and specificity vary with different decision boundaries. The AUC quantifies the overall ability of the model to distinguish between classes [26].

4. Results

The results of the current study demonstrate the effectiveness of the ML techniques in accurately predicting PCOS. Key performance metrics, including accuracy, precision, recall, F1-Score and ROC AUC, were evaluated to assess model reliability. As provided in Table 3.

Table 3: Performance Comparison Between Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC AUC (%)
NB	90.8	82.4	87.5	84.8	96.7
LR	91.7	96.0	85.0	84.2	96.8
SVM	89.9	92.0	90.0	80.7	96.0
RF	89.0	83.3	78.1	80.6	95.0
GBC	89.0	83.3	78.1	80.6	92.1
AdaBoost	88.1	85.2	71.9	78.0	93.4
MLP	87.2	82.1	71.9	76.7	92.1

In terms of accuracy, LR achieved the highest score at 91.7%, indicating its strong overall capability to correctly classify both positive and negative cases. NB followed closely with 90.8%, while SVM and RF achieved 89.9% and 89%, respectively. GBC also matched RF with 89%, and AdaBoost recorded a slightly lower accuracy at 88.1%. The MLP had the lowest accuracy among all models at 87.2%, suggesting it may be less effective in general classification performance for this dataset as shown in Figure 3.

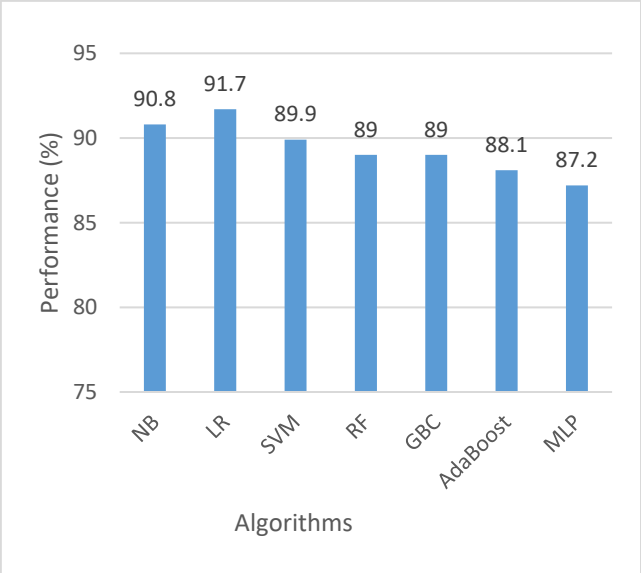

Figure.3: Accuracy Plot of Proposed Models

When evaluating precision, which measures the correctness of positive predictions, LR outperformed all other models with a precision of 96%. SVM came next with 92%, indicating its reliability in predicting relevant positive cases. AdaBoost followed with 85.2%, and both RF and GBC scored 83.3%. NB had a precision of 82.4%, and MLP was the lowest at 82.1%. This metric highlights LR as the most dependable model when minimizing false positives is important as shown in Figure 4.

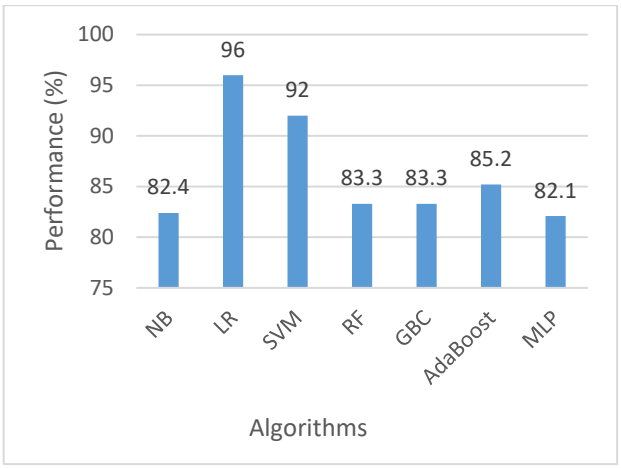

Figure.4: Precision Plot of Proposed Models

The performance comparison based on recall shows that NB achieved the highest recall at 87.5%, demonstrating superior sensitivity in correctly identifying positive cases. This is followed by LR, which also performed well with a recall of 85%, indicating its effectiveness with the dataset's linear characteristics. Meanwhile, SVM, AdaBoost, and MLP exhibited moderate recall values of 79%, reflecting balanced but less outstanding performance in detecting positive cases. Finally, RF and GBC recorded the lowest recall values at 78.1%, suggesting that these ensemble methods may have underperformed in this specific context, possibly due to data characteristics or parameter tuning limitations as shown in Figure 5.

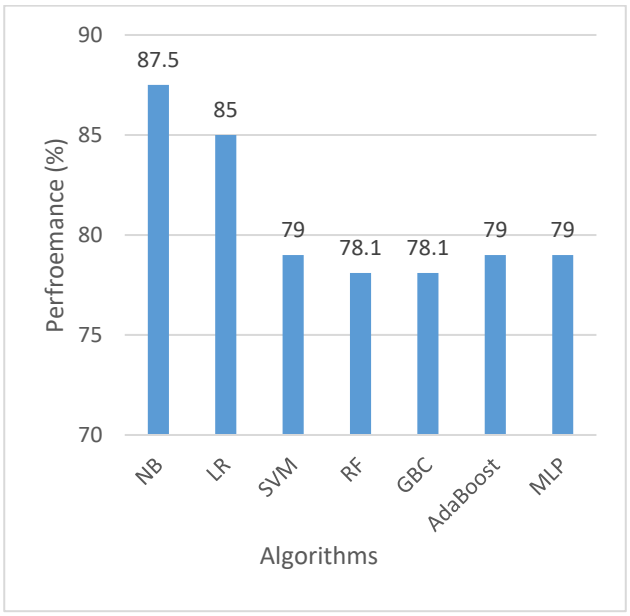

Figure.5: Recall Plot of Proposed Models

For F1 Score, which balances both precision and recall, NB again emerged as the top performer with an F1 Score of 84.8%, suggesting it offers the most balanced predictions. LR was a close second at 84.2%. SVM, RF, and GBC showed similar F1 scores around 80.6–80.7%, reflecting solid but slightly less balanced performance. AdaBoost scored 80%, while MLP had the lowest F1 Score at 76.7%, further confirming its relatively weaker balance

between identifying and correctly classifying positive cases as shown in Figure 6.

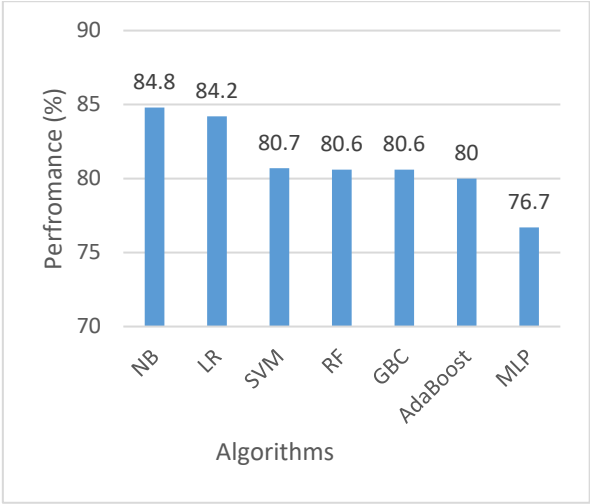

Figure.6: F1 Score Plot of Proposed Models

Regarding ROC AUC, which assesses a model's ability to distinguish between classes at various threshold levels, LR achieved the highest score of 96.8%, closely followed by NB at 96.7% and SVM at 96%. RF also performed well with 95%, and AdaBoost came next at 93.4%. The lowest AUC scores were observed in GBC and MLP, both at 92.1%. These results indicate that while all models demonstrated good class-separating ability, LR and NB were the most effective in this regard as shown in Figure 7.

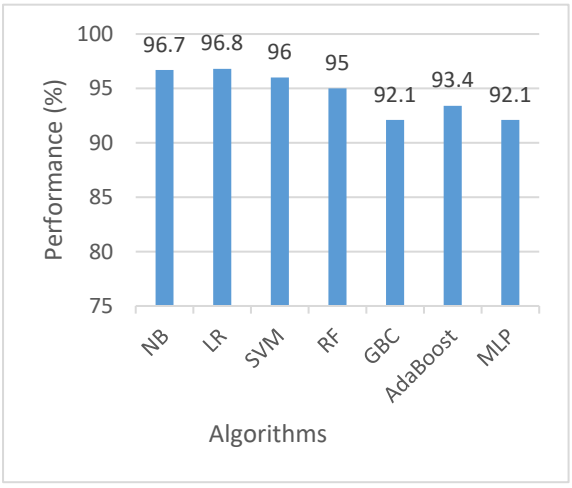

Figure.7: ROC AUC Plot of Proposed Models

5. Discussion

The superior performance of the LR model, achieving the highest AUC, aligns with findings from previous studies discussed in the literature. Similar to the work of Hosain et al. [6], where LR achieved an accuracy of 85.3% due to its strong predictive capability with hormonal and metabolic attributes, this study also demonstrated the effectiveness of LR when supported by appropriate feature selection and data balancing techniques. In the present analysis, class imbalance was effectively managed using the SMOTE algorithm, enhancing the model's sensitivity and specificity—an approach also highlighted by Shanmugavadivel et al. [15] in addressing rare class detection.

Additionally, feature selection using ANOVA F-scores helped identify the most statistically significant predictors, allowing LR to focus on the most influential clinical variables, consistent with the methodology applied by Hosain et al. [6]. These results further validate the literature's emphasis on the importance of simple, interpretable models like LR, particularly when combined with effective preprocessing strategies, achieving performance comparable to or even surpassing more complex models such as RF and SVM [5], [13].

Although the models, particularly LR, achieved high accuracy and AUC scores, we acknowledge that recall values were modest in several cases, indicating a proportion of PCOS cases were not successfully identified. This raises clinical concerns, as missed diagnoses in screening settings may delay treatment. To address this, we will conduct further analysis of false negative cases to identify potential patterns or limitations in feature representation. Additionally, we plan to experiment with threshold tuning, cost-sensitive learning, and advanced resampling methods to improve recall. In clinical contexts, high recall is essential to ensure at-risk patients are not overlooked. A comparative benchmark with clinical diagnostic rates among physicians will also be considered in future work to contextualize the model's performance

6. Conclusion a Future Direction

This study evaluated the performance of seven supervised ML algorithms— LR, NB, SVM, RF, GBC, AdaBoost, and MLP —for the classification of PCOS based on clinical and lifestyle data. The models were assessed using key performance metrics including accuracy, precision, recall, F1 score, and ROC AUC. Among all the models, LR consistently demonstrated the best overall performance.

LR achieved the highest accuracy (91.7%), precision (96%), and ROC AUC (96.8%), and maintained a strong balance between recall and F1 score. Its superior performance can be attributed to the linear separability of the dataset and the model's inherent ability to generalize well with limited assumptions and minimal overfitting. Furthermore, LR is computationally efficient, easy to interpret, and performs reliably when the relationship between features and output is approximately linear characteristics that align well with the nature of this dataset.

This study confirms the potential of machine learning (ML) in identifying PCOS with high accuracy and interpretability. However, limitations such as moderate recall scores, missing hormonal and demographic

variables, and the absence of comparison with clinical decision-making indicate that the current approach requires further enhancement before clinical adoption. Addressing these gaps will improve both the diagnostic value and real-world applicability of ML models in women's health.

Future work should focus on incorporating more comprehensive clinical and biochemical indicators, including insulin resistance markers, androgen levels, and family history. Advanced ensemble techniques like XGBoost and model stacking could be employed to boost predictive performance. Additionally, combining structured data with medical imaging or exploring deep learning (DL) models may lead to more robust diagnostic tools. Expanding the dataset to include diverse populations and validating findings in clinical settings will also be key to ensuring generalizability and fairness in AI-assisted PCOS diagnosis.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The Authors hereby acknowledge that the funding of this paperwork was done and shared across all Authors concerned.

References

[1] C. C. Dennett and J. Simon, "The role of polycystic ovary syndrome in reproductive and metabolic health: overview and approaches for treatment," Diabetes Spectrum,vol. 28, no. 2, pp. 116-120, 2015. DOI: 10.2337/diaspect.28.2.116

[2] I. T. Lee et al., "Depression, anxiety, and risk of metabolic syndrome in women with polycystic ovary syndrome: a longitudinal study," The Journal of Clinical Endocrinology Metabolism,vol. 110, no.3, pp. e750-e756, 2025. DOI: 10.1210/clinem/dgae256

[3] Z. Zad et al., "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," Frontiers in Endocrinology, vol. 15, p. 1298628, 2024. DOI: 10.3389/fendo.2024.1298628

[4] C. Tong, Y. Wu, Z. Zhuang, and Y. Yu, "A diagnostic model for polycystic ovary syndrome based on machine learning," Scientific Reports,vol. 15, no. 1, p. 9821, 2025. DOI: 10.1038/s41598-025-92630-4

[5] P. Chauhan, P. Patil, N. Rane, P. Raundale, and H. Kanakia, "Comparative analysis of machine learning algorithms for prediction of pcos," in 2021 international conference on communication information and computing technology (ICCICT), 2021, pp. 1-7: IEEE. DOI: 10.1109/ICCICT50803.2021.9509757

[6] A. S. Hosain, M. H. K. Mehedi, and I. E. Kabir, "Pconet: A convolutional neural network architecture to detect polycystic ovary syndrome (pcos) from ovarian ultrasound images," in 2022 International Conference on Engineering and Emerging Technologies (ICEET), 2022, pp. 1-6: IEEE. DOI: 10.1109/ICEET56468.2022.10012345

[7] D. Rao, R. R. Dayma, and S. K. Pendekanti, "Deep learning model for diagnosing polycystic ovary syndrome using a comprehensive dataset from Kerala hospitals," International Journal of Electrical Computer Engineering.vol. 14, no. 5, 2024. DOI: 10.11591/ijece.v14i5.36503

[8] B. Panjwani, J. Yadav, V. Mohan, N. Agarwal, and S. Agarwal, "Optimized Machine Learning for the Early Detection of Polycystic Ovary Syndrome in Women," Sensors.vol. 25, no. 4, p. 1166, 2025. DOI: 10.3390/s25041166

[9] L. Ji et al., "Performance of a Full-Coverage Cervical Cancer Screening Program Using on an Artificial Intelligence–and Cloud-Based Diagnostic System: Observational Study of an Ultralarge Population," Journal of Medical Internet Research.vol. 26, p. e51477, 2024. DOI: 10.2196/51477

[10] M. Arya, "Automated detection of acute leukemia using K-means clustering algorithm," 2019. DOI: 10.5120/ijca2019918801

[11] H. O. Boll et al., "Graph neural networks for clinical risk prediction based on electronic health records: A survey," J. Biomed. Informatics.vol. 151, p. 104616, 2024. DOI: 10.1016/j.jbi.2024.104616

[12] M. de Oliveira Gomes, J. de Oliveira Gomes, L. F. Ananias, L. A. Lombardi, F. S. da Silva, and A. P. Espindula, "ANTI-MÜLLERIAN HORMONE AS A DIAGNOSTIC MARKER OF POLYCYSTIC OVARY SYNDROME: A SYSTEMATIC REVIEW WITH META-ANALYSIS," American Journal of Obstetrics Gynecology,2025. DOI: 10.1016/j.ajog.2025.03.077

[13] M. Wang et al., "Biochemical classification diagnosis of polycystic ovary syndrome based on serum steroid hormones," The Journal of Steroid Biochemistry Molecular Biology,vol. 245, p. 106626, 2025. DOI: 10.1016/j.jsbmb.2024.106626

[14] K. M. Mohi Uddin, M. T. A. Bhuiyan, M. M. Rahman, M. M. Islam, and M. A. Uddin, "Early PCOS Detection: A Comparative Analysis of Traditional and Ensemble Machine Learning Models With Advanced Feature Selection," Engineering Reports,vol. 7, no. 2, p. e70008, 2025. DOI: 10.1002/eng2.70008

[15] K. Shanmugavadivel, M. D. MS, M. TR, T. Al-Shehari, N. A. Alsadhan, and T. E. Yimer, "Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data," BMC Medical Informatics Decision Making, vol. 24, no. 1, p. 281, 2024. DOI: 10.1186/s12911-024-02688-9

[16] T. Zohrabi, A. Nadjarzadeh, S. Jambarsang, M. H. Sheikhha, A. Aflatoonian, and H. Mozaffari-Khosravi, "Effect of dietary approaches to stop hypertension and curcumin co-administration on glycemic parameters in polycystic ovary syndrome: An RCT," International Journal of Reproductive BioMedicine,vol. 22, no. 9, p. 689, 2024. DOI: 10.18502/ijrm.v22i9.14994

[17] R. Ahmad, L. A. Maghrabi, I. A. Khaja, L. A. Maghrabi, and M. Ahmad, "SMOTE-Based Automated PCOS Prediction Using Lightweight Deep Learning Models," Diagnostics, vol. 14, no. 19, p. 2225, 2024. DOI: 10.3390/diagnostics14192225

[18] F. Ahmad Musleh, "A Comprehensive Comparative Study of Machine Learning Algorithms for Water Potability Classification," International Journal of Computing Digital Systems,vol. 15, no. 1, pp. 1189-1200, 2024. DOI: 10.12785/ijcds/150112

[19] P. Kottarathil, "Polycystic Ovary Syndrome (PCOS)," https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos, May 10, 2025 2020. DOI: 10.34740/KAGGLE/DSV/1203444

[20] S. Alshakrani, R. Taha, and N. Hewahi, "Chronic kidney disease classification using machine learning classifiers," in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021, pp. 516-519: IEEE. DOI: 10.1109/3ICT51146.2021.9589834

[21] F. A. Musleh and R. G. Taha, "Forecasting of forest fires using machine learning techniques: a comparative study," 2022.

[22] R. Taha, S. Alshakrani, and N. Hewahi, "Exploring Machine Learning Classifiers for Medical Datasets," in 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 255-259: IEEE. DOI: 10.1109/ICDABI53623.2021.9655862

[23] F. Musleh, R. Taha, and A. R. Musleh, "Comparative Analysis of Machine Learning Techniques for Concrete Compressive Strength Prediction," in 2023 4th International Conference on Data Analytics for Business and Industry (ICDABI), 2023, pp. 146-151: IEEE. DOI: 10.1109/ICDABI60145.2023.10629479

[24] E. Dritsas and M. Trigka, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," Computation,vol. 11, no. 2, p. 16, 2023. DOI: 10.3390/computation11020016

[25] X. Hu, A. Yadav, A. Khan, A. P. Sah, and S. Azam, "Construction of PCOS Prediction Model Based on BP Neural Network," in 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2025, pp. 885-889: IEEE. DOI: 10.1109/ICMCSI64620.2025.10883524

[26] A. M. Zeki, R. Taha, and S. Alshakrani, "Developing a predictive model for diabetes using data mining techniques," in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 2021, pp. 24-28: IEEE. DOI: 10.1109/3ICT53449.2021.9582114

Mrs. Ranyah Taha completed her MSc in Big Data Science and Analytics in 2022 through a joint program between Liverpool John Moores University and the University of Bahrain. She earned her BSc in Computer Science from the University of Bahrain in 2018. Her research focuses on leveraging Data Science and Analytics, particularly Machine Learning and Deep Learning, to build advanced models and extract valuable insights from complex datasets. She has contributed to many research papers and was awarded the NASA International Space Apps Challenge – Space Apps Bahrain 2023 Local Impact Award.

Miss Huda Zain El Abdin completed her MSc in Data Science with distinction in 2025. Her graduation project was recognised as one of the top five best graduation projects of the year. She earned her BSc in Software Engineering from the University of Bahrain in 2021. Her research interests lie in the development and application of advanced Natural Language Processing techniques, including LLMs, to solve real-world language understanding challenges. She is particularly interested in the intersection of machine learning and the medical field, exploring how AI can enhance healthcare delivery and diagnostics. She was also awarded second place in a NLP Hackathon organised by London Business School and Middlesex University.

Miss Tala Musleh Pharmacy student at the University of Bahrain, with a keen interest in applying Artificial Intelligence (AI) and Machine Learning (ML) to advance research and clinical practices in the medical and pharmaceutical field.