JOURNAL OF ENGINEERING RESEARCH & SCIENCES

Special Issue

Computing, Engineering and Sciences

March 2025

www.jenrs.com
ISSN: 2831-4085

JENRS

EDITORIAL BOARD
(Special Issue)

Guest Editor

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

Editorial

The dynamic interplay between computing, engineering, and scientific research continues to redefine the boundaries of innovation and discovery. It is with great satisfaction that we present this special issue of the *Journal of Engineering Research and Sciences*, dedicated to *Computing, Engineering and Sciences*. This issue assembles a diverse range of scholarly contributions that collectively emphasize the transformative role of computational technologies in enhancing engineering practices and advancing scientific knowledge.

In recent years, the integration of computing into engineering and scientific domains has become indispensable. The rapid evolution of technologies such as artificial intelligence, machine learning, big data analytics, and high-performance computing has significantly expanded the capabilities of researchers and practitioners alike. The articles featured in this issue reflect how these advancements are being leveraged to address complex challenges, optimize systems, and enable data-driven innovation. By bridging theoretical foundations with practical applications, computing has emerged as a critical enabler of progress across disciplines.

A prominent theme throughout this special issue is the growing importance of data-centric approaches. The ability to collect, process, and analyze vast amounts of data has revolutionized both engineering and scientific methodologies. From predictive maintenance in industrial systems to pattern recognition in scientific datasets, data-driven techniques are enhancing accuracy, efficiency, and reliability. The research contributions included here demonstrate how advanced algorithms and intelligent systems are being applied to extract meaningful insights and support informed decision-making.

Equally significant is the role of modeling and simulation in modern research. Computational simulations allow for the exploration of complex phenomena in controlled and cost-effective environments. This capability is particularly valuable in areas where experimental approaches may be limited by time, resources, or feasibility. The studies presented in this issue showcase how simulation tools are being used to design, analyze, and optimize systems across a wide spectrum of applications, including materials engineering, fluid mechanics, energy systems, and environmental studies.

The advancement of intelligent and automated systems also features prominently in this collection. The integration of embedded systems, robotics, and smart technologies is driving innovation in sectors such as manufacturing, healthcare, transportation, and urban development. These technologies enable the creation of systems that are not only efficient but also adaptive and responsive to changing conditions. The contributions in this issue highlight the potential of such systems to improve productivity, enhance safety, and support sustainable development.

Furthermore, this special issue underscores the importance of interdisciplinary collaboration. The convergence of computing, engineering, and sciences fosters an environment where diverse expertise can be combined to address multifaceted problems. Such collaboration enhances the depth and scope of research, leading to more robust and comprehensive solutions. The works presented here exemplify how interdisciplinary approaches can yield innovative outcomes that extend beyond the capabilities of individual fields.

The editorial team expresses its sincere gratitude to all authors for their valuable contributions and to the reviewers for their careful and rigorous assessments. Their dedication has been instrumental in maintaining the high standards of the journal and ensuring the relevance and quality of this special issue.

As we present this issue, we hope it will serve as both a reflection of current advancements and a source of inspiration for future research. The continued integration of computing with engineering and scientific disciplines holds immense potential for addressing global challenges and fostering sustainable progress. We encourage readers to engage with the research presented herein and to contribute to the ongoing evolution of this vibrant and rapidly advancing field.

Guest Editor

Prof. Paul Andrew

JOURNAL OF ENGINEERING RESEARCH AND SCIENCES

Special Issue **March 2025**

CONTENTS

Received: 14 January, 2025, Revised: 15 February, 2024, Accepted: 16 February, 2025, Online: 31 March, 2025

DOI: https://doi.org/10.55708/js0403002

Software Development and Application for Sound Wave Analysis

Eunsung Jekal [1,*], Juhyun Ku[2], Hyoeun Park[2,3]

[1] Jekal's laboratory, Republic of Korea

[2] Ulsan Pianist Club, Republic of Korea

[3] Scala music institute, Republic of Korea

*Corresponding author: Eunsung Jekal, Jekal's laboratory, Republic of Korea, everjekal@gmail.com

ABSTRACT: In this paper, we developed our own software that can analyze piano performance by using short-time Fourier transform, non-negative matrix decomposition, and root mean square. Additionally, we provided results that reflected the characteristics and signal analysis of various performers for the reliability of the developed software. The software was coded through Python, and it actively utilized Fourier transform to enable precise determination of the information needed to perform a performer's music, such as touch power, speed, and pedals. In conclusion, it shows the possibility that musical flow and waveform analysis can be visually interpreted in a variety of ways. Based on this, we were able to derive an additional approach suitable for designing the system to seamlessly connect hearing and vision.

KEYWORDS: Short-time Fourier transform, non-negative matrix decomposition, music, piano

1. Introduction

Sound wave analysis is essential for understanding music because it contains very complex structures and patterns in terms of science and technology, not just sensory feelings. Sound wave analysis allows a deep understanding of the components of music, which can be used for music creation, learning, and research [1–3]. In fact, acoustic analysis tools have long evolved to enable humans to better understand and utilize sound. These tools used to be more than just an auditory evaluation, but now they have evolved into advanced systems that utilize precise digital signal processing technology [4].

As mentioned above, in the early days of the development of acoustic analysis tools, subjective evaluation using human hearing was dominated. Oscillograph or analog spectrum analyzer allowed most mechanical devices to visually analyze the waveform or frequency of sound waves [5–7]. However, with the advent of the digital age, we began to analyze sound data using computers and software. Fast Fourier Transform (FFT) has become a key technique for frequency domain analysis, and improved precision has enabled accurate temporal and frequency characteristics of sounds [8,9]. With the recent development of machine learning and AI-based analysis, technology that recognizes sound patterns, separates speech and instruments, and analyzes emotions is also being utilized. Real-time sound analysis is possible through mobile devices and cloud computing, and 3D analysis considering spatial sound information is also possible thanks to 3D sound analysis. Extending this to other applications is expected to lead to infinite pioneering in a variety of fields, including healthcare (hearing testing), security (acoustic-based authentication), and environmental monitoring (noise measurement). But there are also obvious limitations.

First, there is a technical limitation of poor analysis accuracy in complex environments. It is difficult to extract or analyze specific sounds in noisy or resonant environments. In addition, it is difficult to process various sound sources, making it difficult to accurately separate sound sources with various characteristics such as musical instruments, human voices, and natural sounds [10].

Another limitation is the lack of practicality. While there are many tools optimized for a particular domain, no general-purpose system has been built to handle all the acoustic data.

Finally, the biggest limitation is the gap between the measurement of analytical tools and the actual hearing of

humans. The difficulty of quantifying subjective sound quality assessments makes it difficult to fully quantify or replace a person's subjective listening experience, and techniques for exquisitely analyzing human cognitive responses (emotions, concentration, etc.) to sound are still in its infancy [11].

2. Ease of Use Necessity needs for analytical tools

2.1. Understanding the Basic Components of Music

Music consists of physical elements such as amplitude, frequency, and temporal structure (rhythm). Sound wave analysis allows for quantitative measurement and understanding of these elements.

Frequency analysis makes it easier to check the pitch of a note and to understand the composition of chords and melodies. Time analysis can identify rhythm patterns and beat structures, and spectrum analysis can identify the unique tone (sound color) of an instrument [12].

2.2. Instruments and Timbre Analysis

Each instrument has its own sound (pitched tone), which comes from the ratio of its fundamental frequency to its background tone. Sound analysis visualizes these acoustic characteristics and helps to understand the differences between instruments.

For example, even at the same pitch as the violin and piano, the difference in tone is due to a combination of frequency components [13].

2.3. Understanding the Emotional Elements of Music

Music is used as a tool for expressing emotions, and certain frequencies, rhythms, and combinations induce emotional responses.

For example, slow rhythms and low frequencies are mainly used to induce sadness, and fast rhythms and high frequencies are used to induce joy. Sound wave analysis can study these relationships to determine the correlation between emotions and music [14].

2.4. A structural analysis of music

The structure of music is not just an arrangement of sounds, but includes complex patterns such as melody, chord, rhythm, and texture. Sound wave analysis allows us to visualize the structural elements of music. For example, harmonic analysis can help musician understand how chords and chords progression, and by analyzing the melody, musician can check the pitch and rhythm pattern. Additionally, if multiple melodies are played simultaneously, they can understand the interaction of each melody [15].

2.5. Support for music production and mixing

Sound wave analysis is essential for solving technical problems in the music production process.

Musicians adjust the sound volume by frequency band to balance the instruments and remove unwanted sound from recorded sound waves. Sound design also helps analyze and improve sound effects [16].

2.6. Music learning and research

Sound wave analysis provides music learners and researchers with tools to visually understand music theory.

2.7. Improved listening experience

Sound wave analysis can visually identify sound elements that are difficult to hear by the human ear (e.g., ultra-low, ultra-high) and improve the listening experience.

For example, we can find hidden detailed sounds in music through spectrograms, and users can visually see the acoustic complexity of music.

The level of analysis and tools required may vary depending on the purpose of analyzing music, but for whatever reason sonic analysis is a powerful tool to explore the scientific and artistic nature of music beyond just listening [17].

3. Prior studies

3.1. A way of expressing sound

There are many ways in which sound is expressed, but it is mainly explained by physical principles such as vibration, waves, and frequency. Sound is a pressure wave that is transmitted through air (or other medium) as an object vibrates. Let's take a closer look at it. Sound is usually produced by an object vibrating. For example, when a piano keyboard is pressed, the strings vibrate, and the vibrations are transmitted into the air to be recognized as sound. This vibration is caused by an object moving and compressing or expanding air particles. And this sound is transmitted through a medium (air, water, metal, etc.). The particles of the medium vibrate, compress and expand to each other, and sound waves are transmitted. At this time, the important concepts are pressure waves and repetitive vibrations.

Compression is a phenomenon in which the particles of the medium get close to each other, and re-action is a phenomenon in which the particles of the medium get away, and the sound propagates by repeating these two processes, through which we hear the sound. These sounds can be distinguished by many characteristics. Mainly, the following factors play an important role in defining sounds.

1) Frequency: The frequency represents the number of vibrations of the sound. At this time, the frequency is measured in Hertz (Hz). For example, 440 vibrations per second are 440 Hz. The higher the frequency, the higher the pitch, and the lower the pitch, the lower the pitch. Human ears can usually hear sounds ranging from 20 Hz to 20,000 Hz.

2) Amplitude: The amplitude represents the volume of the sound, the larger the amplitude, the louder the sound, and the smaller the amplitude, the smaller the sound. Amplitude is an important determinant of the "strength" of sound waves. A larger amplitude makes the sound louder, and a smaller amplitude makes it sound weaker.

3) Wavelength: The wavelength is the distance that a vibration in a cycle occupies in space. The longer the wavelength, the lower the frequency, and the shorter the frequency, the higher the frequency.

4) Timbre (Timbre): Timbre is a unique characteristic of sound, formed by combining various elements in addition to frequency and amplitude. For example, the reason why the piano and violin make different sounds even if they play the same note is that each instrument has different tones. As in this study, in order to analyze sound with software, sound must be digitally expressed, and when expressing sound digitally, the sound is converted into binary number and stored. Digital sound samples analog signals at regular intervals, converts the sample values into numbers, and stores them, which are used to store and reproduce sounds on computers. In summary, sound is essentially a physical vibration and a wave that propagates through a medium. There are two main ways of expressing this, analog and digital, and each method produces a variety of sounds by combining the frequency, amplitude, wavelength, and tone of sound [18–20].

3.2. Traditional method of sound analysis

Frequency analysis is a method of identifying the characteristics of a sound by decomposing the frequency components of the sound. It mainly uses Fourier Transform techniques. Fourier transform is a mathematical method of decomposing a complex waveform into several simple frequency components (sine waves). This transform allows us to know the different frequencies that sound contains.

These Fourier Series and Fourier Transform allow us to analyze sound waves in the frequency domain.

Secondly, spectral analysis is a visual representation of the frequency components obtained through Fourier transform. This analysis visually shows the frequency and intensity of sound.

A spectrogram is a graph that shows the change in frequency components over time, and can visually analyze how sound changes over time.

In addition, time analysis is a method of analyzing sound waveforms over time. This method can track changes in the amplitude of sound over time.

Analyzing the waveform analysis at this time allows users to determine the sound volume, temporal change, and occurrence of specific events.

The waveform is a linear representation of the temporal variation of an analog signal or digital signal, and amplitude and periodicity can be observed.

Users can also track the volume change by analyzing the amplitude of a sound over time. For example, users can determine the beginning and end of a specific sound, or users can analyze the state of attenuation and amplification of the sound.

Further in waveform analysis, characteristic waveform characteristics can also be extracted, which is particularly important for classification or characterization of acoustic signals [21].

3.3. The characteristics of piano sound

The piano is a system with a built-in hammer corresponding to each key, and when the key is pressed, the hammer knocks on the string to produce a sound. The length, thickness, and tension of the string, and the size and material of the hammer are the main factors that determine the tone of the piano. Each note is converted into sound through vibrations with specific frequencies. Frequency is an important factor in determining the pitch of a note. Piano notes range from 20 Hz to 4,000 Hz. They range from the lowest note of the piano, A0 (27.5 Hz), to the highest note, C8 (4,186 Hz).

The pitch is directly related to the frequency, and the higher the frequency, the higher the pitch, and the lower the frequency, the lower the pitch.

The piano's scale consists of 12 scales, separated by octaves. For example, the A4 is 440 Hz, and the A5 doubles its frequency to 880 Hz.

Tone is an element that makes sounds different even at the same frequency. In other words, it can be said to be the "unique color" of a sound.

The tone generated by the piano is largely determined by its structure of harmonics. Since the piano can produce non-sinusoidal waveform sounds, each note contains several harmonics in addition to the fundamental frequencies. This pattern of tone makes the piano's tone unique.

For example, the mid-range of the piano has a soft and warm tone, the high-pitched range has clear and sharp

characteristics, and the low-pitched range has deep and strong characteristics.

Dynamic is the intensity of a sound, or the volume of a sound. In a piano, the volume varies depending on the intensity of pressing the keyboard. The piano is an instrument that allows users to delicately control decremental and incremental dynamics. For example, the piano (p) is a weak sound, and the forte (f) is a strong sound. In addition to this, medium-intensity expressions such as mezzoforte (mf) and mezzo piano (mp) are possible.

In addition, the sound of the piano depends on the temporal characteristics such as attack, duration, and attenuation. Attack is a rapid change in the moment a note begins. The piano's note begins very quickly, and the volume is determined when the hammer hits the string.

The piano's attack is instantaneous and gives a faster reaction than other instruments. For example, the pitch on the piano pops out right away, and the other instruments, such as string or woodwind, can start more smoothly. Duration is a characteristic of how long a sound lasts after it is played. The piano strings gradually decay when they make a sound, because the string's vibration gets weaker and weaker due to friction with the air or other factors. The length, thickness, and tension of the strings affect the duration of the sound at this time. If pianists use a pedal, they can increase the duration of the sound, but when they press the damper pedal, the strings continue to ring without stopping the vibration, making the sound longer. If users look at the waveform, the piano is an instrument that generates non-sine sound. This is a complex waveform, not a sine wave, and several frequency components are mixed to create a rich tone. The sound waves on the piano are rich in harmonics, so they have various tones and rich characteristics. For example, more low-frequency components are included in the lower register, and high-frequency components are more prominent in the upper register.

The notes generated by the piano can be divided into low, medium, and high notes, each range having the following characteristics.

- Blow (A0 to C4): It contains deep, rich, and strong low-frequency components. For example, Blow C1 has a very low frequency of 32.7 Hz.
- Middle tones (C4 to C5): range similar to the human voice, which is the key range of the piano. The mid tones of the piano have a balanced sound and a warm tone.
- High note (C6 to C8): It has a sharp, clear sound, and a clearer sound is produced at a fast tempo or high note.

In conclusion, the characteristics of the sound produced by the piano are influenced by a combination of several factors, including frequency, tone, attack and duration, dynamic, and attenuation. The piano is an instrument with very rich and complex harmonics, and its sound is characterized by fast attacks and various dynamic controls. In addition, the tone is determined by the characteristics of the strings used and the material of the hammer, which makes the piano sound a unique and distinctive sound.

3.4. Characteristics of classical piano music

Piano classical music usually includes classical and romantic music, and its style has characteristic elements in musical structure, dynamics, emotional expression, and technical techniques.

First, complex chords and colorful tones are important features in piano classical music. Since the piano can play multiple notes at the same time, it is excellent at expressing different chords.

The way chords are created is the synthesis of sound waves, which combine several frequency components to create more complex and rich notes. In this process, each note has its own harmonics, which provides a touching and colorful tone to the music.

The second feature is that it delicately controls the dynamics. It can express dramatic changes and subtle emotions by crossing the piano(p) and the forte(f). It also expresses the rhythm and melody by using various technical techniques such as precise rhythms and arpeggios, trills, and scales.

These techniques require fast and repetitive vibrations, resulting in more complicated waveform fluctuations. Lastly, classical piano music focuses on expressing emotional depth, and deals with epic development and emotional flow. The music delicately utilizes the dynamics and rhythm in expressing dramatic contrast or emotional height. When viewed from the perspective of a scientific wave, the notes of classical piano music are not just sine waves but complex non-sinusoidal waveforms. Because of this, the piano's sound includes harmonics, making it richer and more colorful in tone. This sonic quality is very important in classical music.

1) Complex waveform structure. The waveform consists of a fundamental frequency and multiple overtones. For example, the piano's sound vibrates according to its harmonic series, which means that in addition to the fundamental frequency, the background sounds such as 2x frequency, 3x frequency, and 4x frequency are also present. Piano tones have frequencies higher than the basic notes, and they enhance or distort the characteristics of the basic notes. The tone of a piano is formed by the way these tones are nonlinearly

combined. For example, there are many and strong tones at lower notes, and relatively few and microscopic tones at higher notes.

2) (2) Quick attack and sudden waveform changes. In a piano, notes begin quickly, which is a part called an attack. The moment a note begins, the waveform undergoes a drastic change. For example, the sound pressure on a piano rise very quickly as soon as the hammer hits the string, and then it attenuates rapidly. This causes a drastic change in the waveform. The attack part is very short, producing an abnormal waveform indicating a spike with a fast frequency change. The waveform at this moment takes the form of a sharp peak and then a quick decrease in amplitude.

3) Nonlinearity. The sound of classical piano music has nonlinear characteristics, forming unexpected waveforms through multiple nonlinear interactions, even at the same frequency. For example, the moment a hammer hits a string, complex nonlinear oscillations can occur depending on the hammer's mass, speed, and string tension. This results in a mixed waveform in addition to the fundamental frequency, which forms its own tone.

To sum up, the characteristics of piano classical music are very complex and colorful, ranging from its musical composition to the physical characteristics of the sound. Musically, it is characterized by complex chords and various dynamics, and it deals with emotional expression as important. These musical characteristics are physically revealed through the wave peculiarities of sound—complex waveforms, fast attacks, dynamic frequency changes, etc., and the singularity is well represented by the piano's tonal structure and nonlinear waves. The sound waves of the piano are very rich and complex, which contribute to expressing the emotional depth and emotion that classical music is trying to convey well.

4. Method

4.1. Mathematical techniques

4.1.1. Short-time Fourier transform (STFT)

For a given signal x(t), the short-time fourier transform (STFT) is defined as follows:

$$STFT(t,\omega) = \int_{-\infty}^{\infty} x(\tau)\cdot\omega(\tau - t)\cdot e^{-j\omega\tau}\,d\tau$$

Here, $x(t)$ is time domain signal to analyze, while $\omega(\tau - t)$ defined window function depending on t.

The $x(t)$ is called the window function, and it is used to extract only certain sections of the signal. Typical window functions include Hamming, Hanning, and Gaussian windows. The shorter the length of the window function, the higher the time resolution, and the longer the frequency resolution. The sliding window analyzes the signal by sliding the $\omega(\tau - t)$ over time t. After these processes, the results of STFT are given as complex numbers, amplitude represents the strength of the frequency component, and phase represents the phase of the frequency component.

In this case, the information provided in the form of a plural number includes two. The first is the intensity of the frequency expressed in the $|STFT(t,\omega)|$, and the second is the phase information of the frequency component expressed in the $\angle STFT(t,\omega)$.

However, when actually analyzing a signal on the computer, the signal is discrete, so it is calculated by discretizing the STFT for continuous signals. For discrete signal $x[n]$, the STFT is defined as follows:

$$STFT[m,k] = \sum_{n=-\infty}^{\infty} x[n]\cdot\omega[n-m]\cdot e^{-j\frac{2\pi kn}{N}}$$

where $x[n]$ denotes the discrete signal, and $\omega[n-m]$ is the window function applied in the time m.

Since N represents the length of the window or the size of the FFT, the frequency resolution determination is determined by N.

4.1.2. Autocorrelation

Autocorrelation is an important tool for analyzing the self-similarity of signals, measuring how repetitive or periodic they are over time.

The autocorrelation function $R_x(\tau)$ of the continuous signal $x(t)$ is defined as follows:

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)\cdot x(t+\tau)\,dt$$

Autocorrelation functions can be applied in many ways in sonic analysis, first of all they are excellent for periodicity analysis. For example, an automatic correlation function in which a signal is repeated can estimate the fundamental frequency from a voice signal.

It is also used to distinguish noise from useful components in signals. Noise is generally less correlated in time, while useful signals are highly correlated.

And this study can also measure the signal energy mathematically.

$$R_x(0) = \int_{-\infty}^{\infty} x^2(t)\,dt\ (a\ continuous\ signal)$$

$$R_x[0] = \sum_{n=0}^{N-1} x^2[n]\ (discrete\ signal)$$

Finally, it can be used to calculate the period (e.g., pitch) of a speech signal, or to detect a rhythm in a music signal.

Autocorrelation functions are closely related to Fourier transforms. In particular, according to the Wiener-Hinchin theorem:

$$R_x(\tau) \xleftrightarrow{Fourier\ Transform} |X(\omega)|^2$$

In other words, the autocorrelation function $R_x(\tau)$ of the signal pairs the size square of the Fourier transform $|X(\omega)|^2$ of the signal with the Fourier transform $X(\omega)$. This allows us to quickly compute the automatic correlation function via FFT.

$$R_x[k] = \mathcal{F}^{-1}|\mathcal{F}(x)|^2$$

The automatic correlation function is a very important tool in signal analysis and is used for a variety of purposes, including periodicity detection, noise cancellation, and energy calculation. For discrete signals, it can be calculated quickly using FFT and has a wide range of applications such as acoustic analysis, speech recognition, and bio-signal analysis.

4.1.3. *Non-negative matrix factorization (NMF)*

Non-negative matrix factorization (NMF) is a technique that decomposes a non-negative matrix into two non-negative matrices, and is used in various fields such as data analysis, dimensionality reduction, signal processing, and text mining. In this paper, we will explain this mathematically and provide intuition.

The NMF is based on the process of solving the following optimization problems:

$$min_{W,H}\|X - WH\|_F^2$$

The Frobenius norm represents the square of the Euclidean distance between the matrix X and WH.

NMF optimization is classified as a nonlinear optimization problem due to non-negative constraints. To address this, the following update method is used:

$$H \longleftarrow H \circ \frac{W^T X}{W^T W H}$$

$$W \longleftarrow W \circ \frac{X H^T}{W H H^T}$$

At this point, it ends when the convergence condition (e.g., the change in the Frobenius norm is below the threshold), and at each stage, it solves the least-squares problem with non-negative constraints to maintain the non-negative constraints.

NMF is a simple yet powerful matrix decomposition technique that is very useful for extracting patterns in data and interpreting hidden structures. However, it can be effectively utilized only when users understand the limitations and characteristics of NMF and set the appropriate parameters for the data.

4.2. *Physical Tools*

In this paper, we differentiated and coded the acoustic signal analysis tool by providing a user interface. The tool works modular, allowing users to blend and match different tools to tailor their interfaces and features to specific analysis requirements.

The key idea here is the analysis chain. This research interconnect the signal itself or the signal analysis results to different windows to form a functional block sequence. It includes file input, data collection modules, FFT analysis, measuring instruments, etc.

There is only one segment of the signal needed to visually represent a musical signal. Efficient physical tools were used to effectively imitate real-time signal analysis using pre-recorded offline signals.

The sequence of all values is "signal" at this time. The concept of "signal" includes an array of all X/Y values, including audio signals, spectra, or other data representations. This broad definition means that all sequences are attributed to the sampling rate, which can also be applied to data not derived from digital sampling of analog signals. For example, the "sampling rate" of the FFT analysis results is determined by the number of bins (values) per unit of frequency (Hz) on the X-axis.

This framework enables a flexible analysis approach, such as performing FFT on the results of previous FFTs. The biggest advantage of this work is that there are no restrictions on analytical exploration. I would like to compare and analyze classical piano music while pioneering a unique methodology to achieve the desired insights.

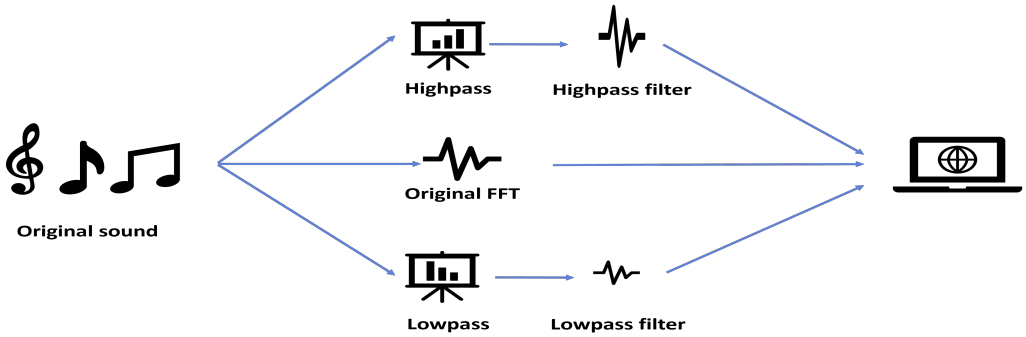

Figure 1: A schematic diagram of the process by which actual music is represented in a graph.

```python
import librosa
import numpy as np
import matplotlib.pyplot as plt

# make signal
sr = 22050  # sampling rate
t = np.linspace(0, 2, sr * 2, endpoint=False)  # 2s signal
x = 0.5 * np.sin(2 * np.pi * 440 * t)  # 440Hz sin wave

# STFT calculation
stft_result = librosa.stft(x, n_fft=2048, hop_length=512, win_length=2048, window='

spectrogram = np.abs(stft_result)

# visualization
plt.figure(figsize=(10, 6))
librosa.display.specshow(librosa.amplitude_to_db(spectrogram, ref=np.max),
                         sr=sr, hop_length=512, x_axis='time', y_axis='log')
plt.colorbar(format='%+2.0f dB')
plt.title('Spectrogram')

```

Figure 2: Python code for STFT

The above code is a Python code that calculates the STFT of a 440Hz sine wave and visualizes it as a spectrogram.

```python
import numpy as np
import matplotlib.pyplot as plt

# make signal
fs = 1000  # sampling frequency
t = np.linspace(0, 1, fs, endpoint=False)  # for 1s
freq = 5  # frequency (Hz)
signal = np.sin(2 * np.pi * freq * t)  # make sin wave

# calculation
def autocorrelation(x):
    n = len(x)
    result = np.correlate(x, x, mode='full')
    return result[result.size // 2:] / n

auto_corr = autocorrelation(signal)

# visualization
lags = np.arange(len(auto_corr))
plt.figure(figsize=(10, 6))
plt.plot(lags, auto_corr)
plt.title("Autocorrelation of Signal")
plt.xlabel("Lag")
plt.ylabel("Autocorrelation")
plt.grid()
plt.show()
```

Figure 3: Python code for calculating the automatic correlation function of the signal.

```python
import numpy as np
from sklearn.decomposition import NMF

# make data
X = np.abs(np.random.randn(10, 8))  # ramdom

# NMF model
k = 3  # dimension
model = NMF(n_components=k, init='random', random_state=42)

# NMF calculation
W = model.fit_transform(X)
H = model.components_

# result
print("Original Matrix (X):")
print(X)
print("\nBasis Matrix (W):")
print(W)
print("\nCoefficient Matrix (H):")
print(H)
print("\nReconstructed Matrix (W * H):")
print(np.dot(W, H))

```

Figure 4: Python code for implementing NMF.

The code above generates a sine wave of 5 Hz and calculates the automatic correlation function of the signal using np.correlate. The visualization of the result is then plotted according to the lag.

When implementing NMF in Python, the Sklearn library makes it easy to handle.

4.3. The specifications of a piano

The piano used for the performance and recording was the Yamaha Grand Piano GC1, which was produced in Japan. There were a total of 88 keys, a soft pedal (left), a Sostenuto pedal (center), and a damper pedal (right). The top lid was open in the recording environment.

The depth of the grand piano varies depending on the model, but it usually ranges from 151 cm to 188 cm. For the grand piano used in this study, it should be noted that the total length from the keyboard to the longest string end is 161 cm, therefore the actual sound and the recorded sound may differ if no sinusoidal waves are made at that length.

4.4. Characteristics of performed music and performers characteristics

4.4.1. Twinkle, Twinkle, Little Star

The original song is titled “Mozart Variations 12 on 'Ah, vous dirai-je, Maman' in Camajor, K265", but it is popularly known as a twinkling little star.

The theme is a folk melody consisting of 12 simple and simple bars. It is composed in C major and features a clear and clear sound pattern.

4.4.2. Gavotte composed by Cornelius Gurlitt

In this paper, we chose Gavotte composed by Cornelius Gurlitt because it is a specific song that can show the connection and disconnection of notes, and changes in articulation, a playing technique that represents legato and staccato, due to its relatively fast rhythm. It also has the advantage of being able to clearly express the visual through the waveform graph because the phrase section on the score is clear.

A total of five notes are connected to the regato until the first note of the next bar, followed by two staccato notes. If we look closely at this part, this study can distinguish the waveform difference between the regato and staccado methods.

There's also a part that gradually becomes a crescendo from the 9th bar to the 12th bar, and from the 13th bar to the 16th bar, this study can even look at the volume in a single piece because it's played quietly with the right hand only without the left hand.

4.4.3. Rachmaninov Piano Concerto No. 2 in c minor, Op. 18

In the case of the first seven bars, the two-handed chord that is pressed at the same time and the left-handed bass F alternate. In addition, since the right-handed inner part

changes with each bar, it is easy to observe the change of sound waves with right-handed chords.

The volume of the sound is also gradually increasing from pp to ff, so this study can see if the "feel of something coming from a distance" is also expressed on the graph when drawn in a waveform.

4.4.4. *Characteristics of performers*

1) Younju Kim, Female: 166cm, 54kg. She is characterized by having relatively long arms and is a performer with good movement. However, instead of having a large arm movement, the force cannot be transmitted quickly to the fingertips, making it difficult to perform strong strokes. The sound resonates well, but the amplitude is not large.
2) Juhyun Ku, Female: 163.5cm, 60kg, hand size: From the first note of Do to the next octave Mi. The weight of the arms is heavier and the fingertips are harder than other pianists. There is less movement of the arms and the fingertips are attached to the piano to control the speed. It is characterized by less movement of the arms.
3) Hyoeun Park, Female : 161cm, 47kg. The weight of the arms is light and the movement is fast. She is one of the musicians who produce accurate sounds. When playing the piano, the movement is big, but the time when the hands are attached to the piano keys is short.
4) Eunsung Jekal, Female: 158, 42kg, beginner. Unlike other pianists, he recorded on the Kawai Digital Piano. He has been learning for about a year, so he has no movement in his arms and is very weak in other health compared to professional pianists.

5. Results

5.1. *Magnitude of sounds*

Root Mean Square (RMS) represents the average energy of a signal and is the basic method for quantitative comparison of sound magnitudes as shown in figure 5. In this paper, two recorded files were imported into the software and calculated as the rms function of MATLAB. Afterwards, this study measured the Loudness Units Full Scale (LUFS). LUFS is an international standard that measures the loudness of sounds based on the volume felt by the human ear. The LUFS values of the two performances were compared using the Loudness Normalization function, and it is generally considered that the lower the LUFS value, the higher the volume. That is, the size increases as it approaches zero. Finally, this research analyzed decibels (dB), and we gave it as fig.6. Decibels compare the loudness based on the peak amplitude or average amplitude of the signal. In this paper, the Peak Level and Average Level were measured using their own software, and the Peak Amplitude was calculated using MATLAB's max (abs(signal).

Performer Jekal:
RMS: -20.5dB
LUFS: -15 LUFS
Dynamic range: 10 dB

Performer Ku:
RMS: -18.2dB
LUFS: -12 LUFS
Dynamic range: 14 dB

In conclusion, Ku makes an overall louder sound, and the dynamic range is also wider, showing that it is more expressive.

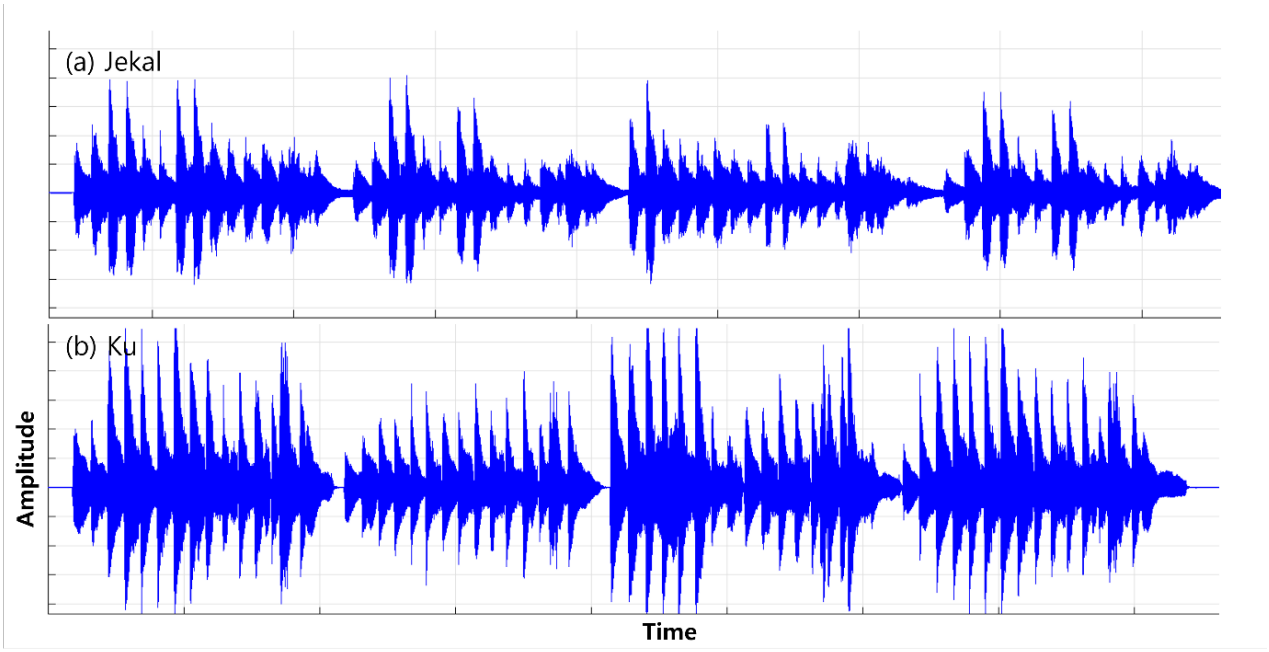

Figure 5: Sound waves of 'Twinkle, Twinkle, Little Star' performed by (a) Jekal and (b) Ku.

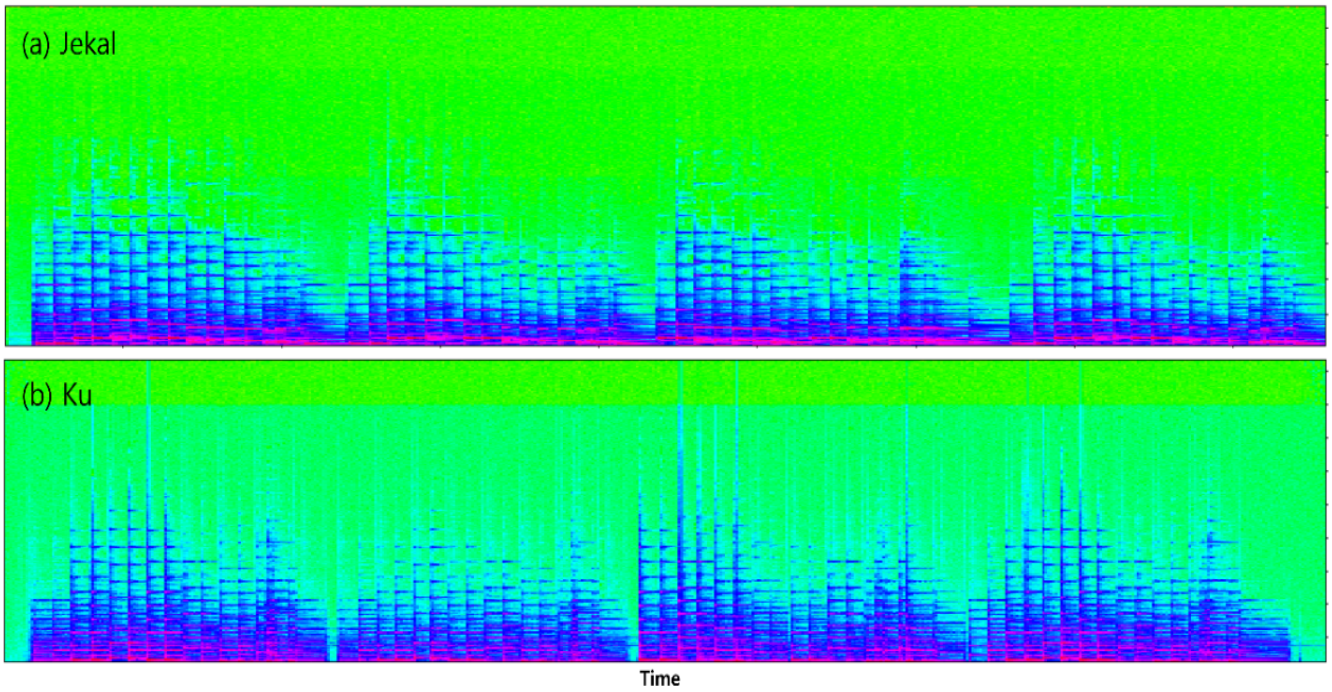

Figure 6: Sound volume of (a) Jekal and (b) Ku

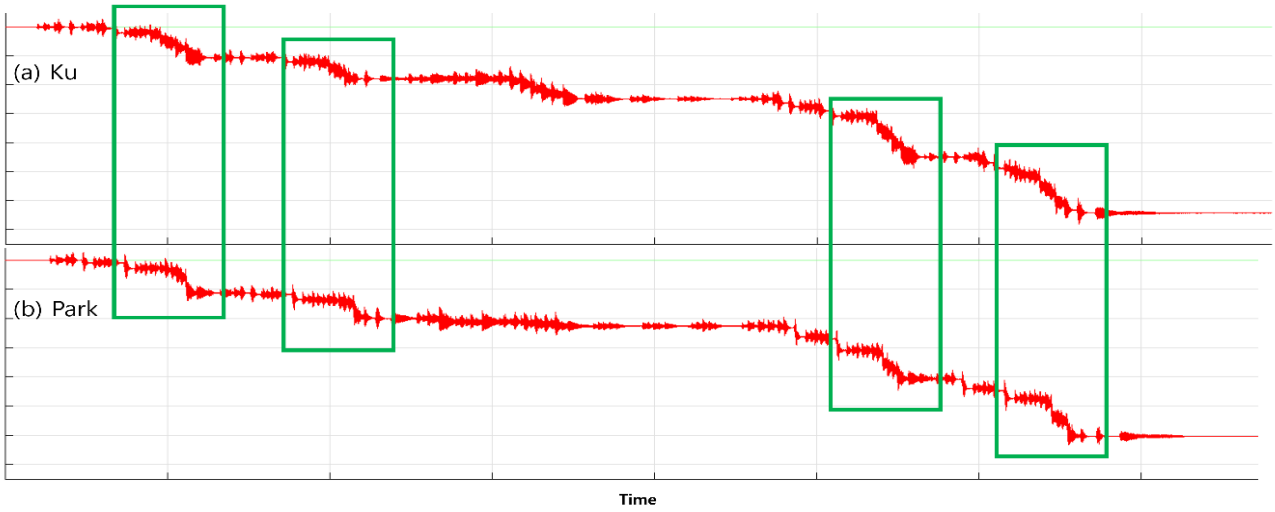

Figure 7: Differential graphs of waveforms for velocity analysis

5.2. Velocity

By differentiating the waveform graph, this research can get information that represents the rate of change of the signal. This rate of change can be interpreted as speed, which varies depending on the physical or mathematical properties of the signal. For acoustic signals, differentiation is useful for analyzing the rate of amplitude change or for deeply understanding the characteristics of the signal.

1st differential expressed as

$$v(t) = \frac{dx(t)}{dt}$$

2nd differential expressed as

$$a(t) = \frac{d^2x(t)}{d^2t}$$

Here, if it is calculated as a discrete signal instead of a continuous signal, it may be approximated as follows:

$$v[n] = \frac{x[n+1] - x[n]}{\Delta t}$$

Figure. 7 shows the change in speed by differentiating the waveform. Comparing (a) and (b) in the boxed sections, this research can see that the change in (b) is more rapid. In other words, (b) the performer hits one note while playing the piano and then moves on to the next note compared to (a). This may have been due to the performer's movement or body shape.

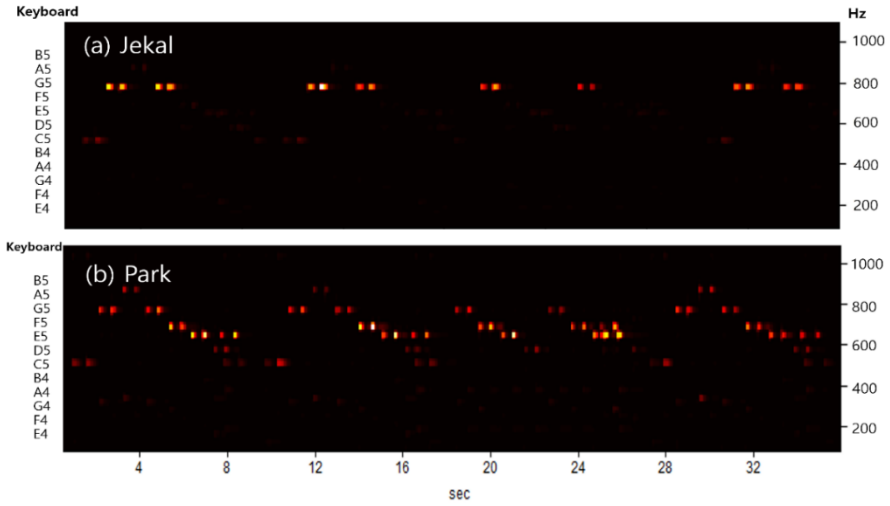

Figure 8: Piano keyboard touch strength of (a) Jekal and (b) Park

5.3. Touch intensity

To mathematically analyze the strength (touch strength) of pressing the keys when playing the piano, users need to obtain and analyze data related to the force on the keys through acoustic signals or physical sensors. This can be represented by the amplitude of an acoustic signal, or the physical movement of the keys.

In this paper, amplitude-based analysis was utilized. This is because the amplitude of the generated acoustic signal increases when the keyboard is pressed hard, so the intensity can be estimated by analyzing the amplitude.

For this purpose, the recorded acoustic signal was imported into its own analysis software, and the average amplitude was measured by calculating the RMS value of the signal.

The RMS is calculated as follows:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x[i]^2}$$

Here, $x[i]$ expresses value of sample signal and N denotes number of samples.

Gavotte composed by Cornelius Gurlit, expressed in the form of sound waves in fig.9, can be largely divided into six parts. Sections 1 and 2, and sections 5 and 6 are all played with the same note and rhythm. However, in section 4, it can be seen that the sudden sound becomes smaller, and the difference between the three pianists can be seen more clearly here. The note played by Kim grows and decreases again, the note of the beat becomes louder and louder, and the note of the sphere becomes smaller and smaller. In the case of part 5 as well, Kim starts with a loud sound following section 4 and Park starts with a small sound before playing louder and louder. In fact, it is often difficult to hear the change in detail when listening to fast songs such as Gavotte. However, the software used in this study shows changes in the overall structure and musical image of the song well through the waveform graph.

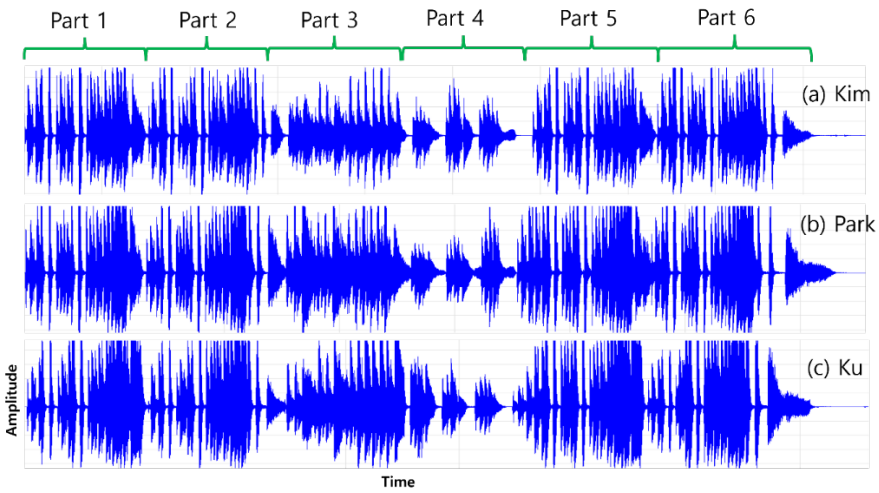

Figure 9: Sound waves of 'Gavotte' performed by (a) Kim, (b) Park and (c) Ku.

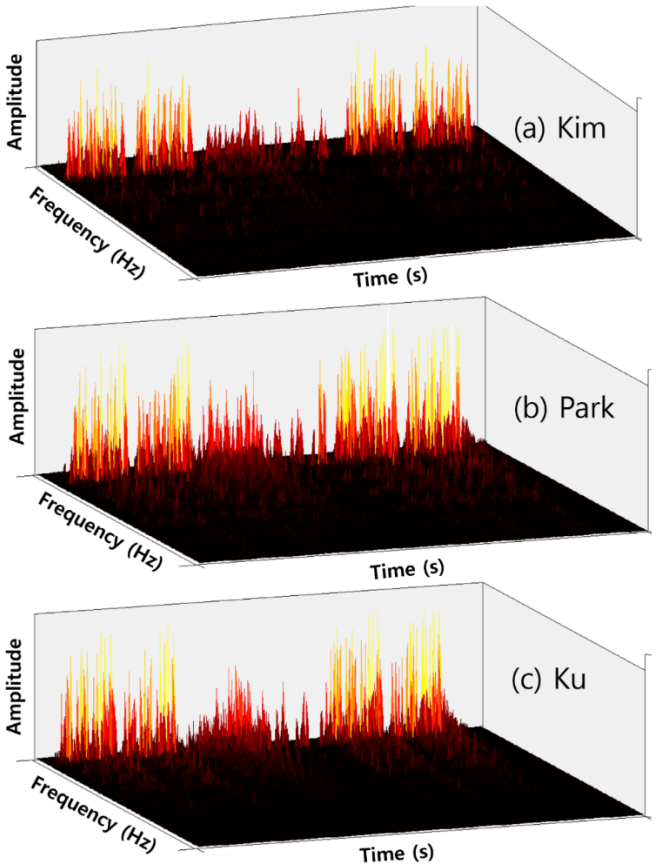

Figure 10: Sound flows of 'Gavotte' performed by (a)Kim, (b)Park and (c)Ku.

5.4. Music flow

Graph theory was used in this paper because it can be very useful in structural and relational analysis of piano music, and the results are shown in Fig.10.

First, the graph G consists of node set V and edge set E, so it is expressed as $G = (V, E)$.

Music has a direction, so this study need a graph that considers the direction here, and the in-degree is expressed as

$$deg_{in}(v) = |\{(u, v) \in E\}|$$

The out-degree is expressed as

$$deg_{out}(v) = |\{(v, u) \in E\}|$$

and the sum of entry and exit orders for all nodes is expressed as

$$\sum_{v \in V} deg_{in}(v) = \sum_{v \in V} deg_{out}(v) = |E|$$

Entropy of graph G denoted as

$$H(G) = -\sum_{v \in V} p(v) log p(v),$$

where $p(v) = \frac{\deg(v)}{2|E|}$.

In this way, graph theory can be used to create a key tool for analyzing music or solving problems that arise during the practice process.

6. Discussion

In this work, we explore ways to improve the smooth connection of visual and auditory flows through waveform analysis of music.

The sound was analyzed using mathematical techniques such as Short-Time Fourier Transform (STFT), Non-Negative Matrix Factorization (NMF), and Root Mean Square (RMS), and the analysis results reflecting the characteristics of various performers were presented. Analyzing sound in this way allows various applications such as emotional expression, structural analysis, understanding differences in musical instrument tones, and supporting music production.

The sound volume analysis compared the size of the performance and the dynamic range through the analysis of the performer's RMS, LUFS, and dB.

For velocity analysis, the rate of change of the signal and the difference in the movement of the performer were evaluated through the first and second derivatives of the waveform. As for the touch intensity, the intensity of the

keyboard touch was estimated by RMS, and the expressive power of the player was compared.

In addition, in the flow of music, the structural change of the song was visualized as a waveform graph to clearly confirm the difference in the interpretation of the performer.

7. Limitations and future works

7.1. Overcome the difference between soundproof room and hall

This graph is a visualized graph of three performances from the introduction of movement 1 to 9 of Rachmaninoff Piano Concerto No. 2. Although it is a concerto, it is easy to compare the sound wave graphs in the soundproof room and hall because only the piano is played at the introduction without an orchestra.

Figure 11(a) and (b) are recorded on the same piano in the same soundproof room, and Fig.11(c) is recorded in a large concert hall.

Looking at the green box, it can be seen that the up-and-down symmetry of the sound wave is greatly broken in Fig.11(c) played in the hall, which is seen as irregularity in the effect on the echo of the hall.

7.2. Improved visual-audible connection smoothness

Music waveforms can naturally have irregular, sharp shapes. To visualize them seamlessly, the signals need to be processed smoothly.

Low-pass Filter can be utilized for smoothing the signal, which reduces the high frequency component (noise) and softens the waveform.

$$y[n] = \frac{1}{N}\sum_{k=0}^{N-1} x[n-k]$$

In addition, Spline Interpolation creates curves that seamlessly connect data points, enabling smoother waveform representations.

There's another way to enhance visual synchronization. It reinforces visual effects, which align with auditory perception, by emphasizing features such as music's rhythm, tempo, and range. For example, users might consider extracting the music's rhythm through beat detection algorithms to synchronize visual "beat" or change the color according to frequency band or amplitude.

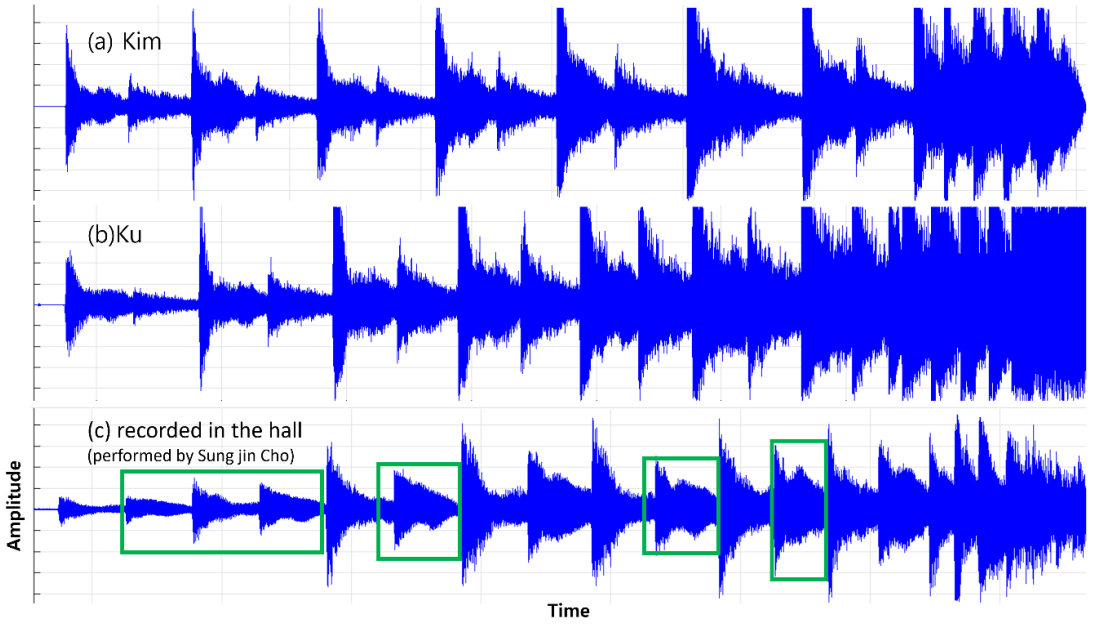

Figure 11: Sound waves of ''Rachmaninoff Piano Concerto No. 2" performed by (a)Kim, (b)Ku and (c)Sung jin Cho.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] D. Arfib, "Digital synthesis of complex spectra by means of multiplication of nonlinear distorted sine waves," 1979.

[2] B. Bank, "Nonlinear Interaction in the Digital Waveguide With the Application to Piano Sound Synthesis.," in ICMC, 2000.

[3] B. Bank, V. Valimaki, "Robust loss filter design for digital waveguide synthesis of string tones," *IEEE Signal Processing Letters*, vol. 10, no. 1, pp. 18–20, 2003, doi:10.1109/LSP.2002.806707.

[4] J. Bensa, S. Bilbao, R. Kronland-Martinet, J.O. Smith III, "The simulation of piano string vibration: From physical models to

finite difference schemes and digital waveguides," *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 1095–1107, 2003, doi:10.1121/1.1587146.

[5] J. Berthaut, M.N. Ichchou, L. Jézéquel, "Piano soundboard: structural behavior, numerical and experimental study in the modal range," *Applied Acoustics*, vol. 64, no. 11, pp. 1113–1136, 2003, doi:https://doi.org/10.1016/S0003-682X(03)00065-3.

[6] G. Borin, G. De Poli, D. Rocchesso, "Elimination of delay-free loops in discrete-time models of nonlinear acoustic systems," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 597–605, 2000, doi:10.1109/89.861380.

[7] C. Cadoz, A. Luciani, J. Florens, C. Roads, F. Chadabe, "Responsive Input Devices and Sound Synthesis by Stimulation of Instrumental Mechanisms: The Cordis System," *Computer Music Journal*, vol. 8, no. 3, pp. 60–73, 1984, doi:10.2307/3679813.

[8] A. Chaigne, A. Askenfelt, "Numerical simulations of piano strings. I. A physical model for a struck string using finite difference methods," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1112–1118, 1994, doi:10.1121/1.408459.

[9] A. Chaigne, A. Askenfelt, "Numerical simulations of piano strings. II. Comparisons with measurements and systematic exploration of some hammer-string parameters," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1631–1640, 1994, doi:10.1121/1.408549.

[10] J.M. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the Audio Engineering Society*, vol. 21, no. 7, pp. 526–534, 1973.

[11] H.A. Conklin Jr., "Piano design factors—Their influence on tone and acoustical performance," *The Journal of the Acoustical Society of America*, vol. 81, no. S1, pp. S60–S60, 2005, doi:10.1121/1.2024314.

[12] A. Fettweis, "Wave digital filters: Theory and practice," *Proceedings of the IEEE*, vol. 74, no. 2, pp. 270–327, 1986, doi:10.1109/PROC.1986.13458.

[13] J.L. Flanagan, R.M. Golden, "Phase Vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966, doi:https://doi.org/10.1002/j.1538-7305.1966.tb01706.x.

[14] H. Fletcher, E.D. Blackham, R. Stratton, "Quality of Piano Tones," *The Journal of the Acoustical Society of America*, vol. 34, no. 6, pp. 749–761, 1962, doi:10.1121/1.1918192.

[15] K. Karplus, A. Strong, "Digital Synthesis of Plucked-String and Drum Timbres," *Computer Music Journal*, vol. 7, no. 2, pp. 43–55, 1983, doi:10.2307/3680062.

[16] J. Laroche, J.-L. Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 329–344, 1994, doi:10.1109/89.279282.

[17] M. Le Brun, "Digital waveshaping synthesis," *Journal of the Audio Engineering Society*, vol. 27, no. 4, pp. 250–266, 1979.

[18] T. Singh, M. Kumari, D.S. Gupta, "Rumor identification and diffusion impact analysis in real-time text stream using deep learning," *The Journal of Supercomputing*, vol. 80, no. 6, pp. 7993–8037, 2024, doi:10.1007/s11227-023-05726-x.

[19] T. singh, M. Kumari, D.S. Gupta, "Context-Based Persuasion Analysis of Sentiment Polarity Disambiguation in Social Media Text Streams," *New Generation Computing*, vol. 42, no. 4, pp. 497–531, 2024, doi:10.1007/s00354-023-00238-x.

[20] T. Singh, M. Kumari, D.S. Gupta, "Real-time event detection and classification in social text steam using embedding," *Cluster Computing*, vol. 25, no. 6, pp. 3799–3817, 2022, doi:10.1007/s10586-022-03610-6.

[21] H. Li, K. Chakraborty, S. Kanemitsu, "Music as Mathematics of Senses," *Advances in Pure Mathematics*, vol. 08, no. 12, pp. 845–862, 2018, doi:10.4236/apm.2018.812052.

Received: 13 January, 2025 Revised: 02 March, 2025 Accepted: 03 March, 2025 Online: 20 May, 2025

DOI: https://doi.org/10.55708/js0405001

Water Potability Prediction Using Neural Networks

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: The crucial need for maintaining specific water potability levels depending on the sector of utilization, this is becoming increasingly challenging due to the increased pollution. It is therefore important to have fast and reliable water potability assessment techniques. A subset of Machine Learning (ML); being Deep Learning (DL), can be utilized to develop models capable of measuring water quality while assessing its potability with high levels of accuracy; thus, ensuring that water meets the set standards based on the required sector of utilization. In this research, the effectiveness of Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) Neural Networks (NN) were contrasted for Water Quality Classification (WQC). The MLP model demonstrated superior performance, achieving higher precision, accuracy, F-measure, recall and the area under the receiver operating characteristic curve (ROC-AUC) scores, indicating its effectiveness in this application compared to the LSTM approach. The experimental findings revealed that MLP NN model outperformed the LSTM NN model in WQC tasks. The MLP model achieved very high performance with an accuracy of 99.9%, an F-measure of 99.9%, a precision of 99.9%, a recall of 99.9%, and a ROC-AUC of 100%, significantly outperforming the LSTM model, which attained an accuracy of 97.6%, an F-measure of 97.1%, a precision of 97.1%, a recall of 97.5%, and a ROC-AUC of 97.9%. The study's novelty lies in employing DL for binary classification, yielding outstanding outcomes in the crucial domain of WQC.

KEYWORDS: Artificial Intelligence, Data Analysis, Water Quality Classification, Neural Networks, Civil Engineering.

1. Introduction

Water is the main and fundamental component in all biological processes; hence, it is the principal element crucial for sustaining all forms of life and a balanced ecosystem. Additionally, being in the industrial age, water became vital in a range of manufacturing processes; moreover, as time goes water is introduced into further domains; therefore, increasing the water demand necessary to sustain all these applications [1].

As this demand for water increases the issue of water scarcity even grows deeper; this is because of the lack of clean, directly utilizable water sources; in addition to, the continued pollution of water sources due to unresponsible human behavior. This dictates water treatment processes that transform contaminated water into untainted water. The overlapping of water into a variety of domains means that there is a spectrum of water quality levels depending on the sector [2].

Water quality and potability are closely related, as evaluating water quality involves assessing its appropriateness based on its physical, chemical, and biological attributes for different uses. Potability specifically focuses on whether water is safe for human consumption, ensuring it meets health standards and is free from harmful contaminants. Water quality refers to the overall status of water as reflected through its biological, chemical, and physical characteristics, determining its suitability for various uses, such as agriculture or industrial processes. The differences between them affect their assessment techniques: water quality assessments may include a broader range of tests for pollutants and environmental impact, while potability focuses on specific contaminants and health-related criteria to ensure drinking safety. Understanding these distinctions is crucial for effective water management and public health [2].

Conventional methods have been employed to evaluate water quality, particularly to ascertain its potability; however, these traditional approaches can now be replaced by new procedures that ensure higher accuracy and precision. These new techniques are equipped with the usage of Artificial Intelligence (AI) and specifically the usage of DL, a branch of ML [3].

Machine learning incorporates the utilization of mathematical models that can identify the trends and characteristics through which water is classified into contaminated or untainted depending on the domain. Thus, through a range of variables, water quality and water potability are instantly assessed and classified with nearly faultless accuracy [3].

This study seeks to assess and compare the potency of two Neural Network (NN) architectures, LSTM and MLPs, for WQC; NNs are the fundamental components of DL. The study will abide by the Cross-Industry Standard Protocol for Data Mining (CRISP-DM) methodology working on a dataset obtained from a government-maintained official website in India.

The upcoming segments of this article will be organized beginning with literature review, followed by methodology, description and preparation of the dataset, moving on to the classification algorithm, the outcomes, and concluding with both the discussion and the conclusion.

2. Literature Review

Due to the significance of the topic, much research has been targeted at the topic of water potability forecasting. This section will highlight some of these contributions to set a scientific base for comparison with the results obtained in this work.

To start, a study conducted by [4] investigates the applicability of AutoDL in Water Quality Assessment. AutoDL is an emerging field, automating DL pipelines; subsequently, comparing its performance against conventional models. Results show that conventional DL outperforms AutoDL by 1.8% for binary class data and 1% for multiclass data, with accuracies ranging from ~96% to ~99%.

Additionally, In [5], the authors carried a study focused on forecasting water's quality in one of Greece's lakes, conventional ML models such as Support Vector Regression (SVR) and Decision Tree (DT) were pitted against DL models like LSTM, Conventional Neural Network (CNN), and a hybrid CNN-LSTM model. The objective was to predict levels of Chlorophyll-a (Chl-a) and Dissolved Oxygen (DO) using physicochemical variables collected between June 2012 and May 2013. The novel merged approach showed improved performance over standalone models. Lag times of up to two intervals were used for prediction. LSTM excelled in DO prediction, while both DL models performed similarly for Chl-a. The merged CNN-LSTM approach demonstrated superior predictive accuracy for both variables, effectively capturing variations in DO concentrations. Evaluation metrics included correlation coefficients, RMSE, MAE, and graphical analyses, revealing the hybrid model's enhanced predictive capabilities in capturing diverse water quality levels.

Moreover, research proposed by [6] addresses the precise estimation of Effluent Total Nitrogen (E-TN) for optimizing the operations of Wastewater Treatment facilities (WWTPs), ensuring regulatory compliance and reducing energy usage. The complexity inherent in WWTPs poses a significant obstacle for accurate multivariate time series forecasting of E-TN due to their intricate nonlinear nature. To tackle this challenge, a new predictive framework is proposed, integrating the Golden Jackal Optimization (GJO) algorithm for feature selection and a hybrid DL architecture, the CNN-LSTM-TCN (CLT) model. CLT combines CNN, LSTM, and TCN to capture complex interrelations within WWTP datasets. A two-step feature selection process enhances prediction precision, with GJO fine-tuning CLT hyperparameters. Findings underscore the effectiveness of the proposed system in precisely forecasting multivariate water quality time series in WWTPs, demonstrating superior performance across diverse prediction scenarios.

In this study performed by [7] supervised learning is utilized to develop precise predictive models from labeled data, aiming to categorize water as safe or unsafe based on its characteristics. Various ML models are assessed for binary classification using features like physical, chemical, and microbial parameters. The findings demonstrate that the Stacking model, in combination with SMOTE and 10-fold cross-validation, it surpasses other methods, yielding remarkable outcomes. Notably, it demonstrates an accuracy and recall rate of 98.1%, precision of 100%, and an AUC value of 99.9%.

Furthermore, a study by [8] introduces a merged IoT and ML system for detailed water quality forecasting. By analyzing Rohri Canal data in SBA, Pakistan, the system predicts Water Quality Class and Water Quality Index (WQI). ML models like LSTM, SVR, MLP, and Nonlinear Autoregressive Neural Network (NARNet) predict WQI, while Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), DT, and Random Forest (RF) forecast WQC. Results indicate that the MLP regression model excels with the lowest errors and highest R-squared (R2) of 0.93. RF leads in classification, achieving high precision (0.93), accuracy (0.91), recall (0.92), and a F1-score of 0.91. Notably, models perform better with smaller datasets. This study demonstrates enhanced regression and classification performance compared to previous research.

In another study conducted by [9] a system combining a Discrete Wavelet Transform (DWT), LSTM, and an Artificial Neural Network (ANN) was created to predict the Jinjiang River's water quality. Initially, a MLP-NN handled missing values within the water quality dataset. Subsequently, the Daubechies 5 (Db5) wavelet divided the data into low and high-frequency signals, serving as LSTM inputs for training and prediction. Comparative analysis against various models such as NAR, Autoregressive Integrated Moving Average (ARIMA), ANN-LSTM, MLP, LSTM, and CNN-LSTM demonstrated the superior performance of the ANN-WT-LSTM model across multiple evaluation n metrics, highlighting its effectiveness in the forecasting of the quality of water.

To evaluate the efficacy of the proposed DL method incorporating LSTM, Gated Recurrent Unit (GRU), and Recurrent Neural Network (RNN) a case study was undertaken in a southern Chinese city by [10]. A comparison was made between the proposed method, a linear approach which is the Multiple Linear Regression, (MLR), and a traditional learning algorithm (MLP). The DL algorithm demonstrated strong predictive capabilities, with GRU outperforming in predicting water quality chemical indices and exhibiting a swifter learning curve. Findings indicated that GRU outperformed traditional ML algorithms by 9.13%–15.03% in terms of R2, surpassed RNN and LSTM by 0.82%–5.07%, and exceeded linear methods by 37.26%–43.38% for the same parameter.

Moving forward, in a study proposed by [11] a novel system for monitoring drinking water potability, emphasizing sustainability and environmental friendliness was introduced. An adaptive neuro-fuzzy inference system (ANFIS) was created for WQI prediction, while K-nearest neighbors (KNN) and Feed-forward neural network (FFNN) were utilized for WQC. The ANFIS model excelled in WQI prediction, while the FFNN algorithm achieved 100% accuracy in WQC. Notably, ANFIS demonstrated 96.17% accuracy in WQI prediction during testing, while FFNN remained robust in classifying water quality.

In the research led by [12], sophisticated AI algorithms were formulated to forecast the Water Quality Index (WQI) and categorize water quality. SVM, K-NN, and Naive Bayes algorithms were utilized for Water Quality Classification (WQC) forecasting while NARNet and LSTM DL models were applied for WQI prediction. The assessment of the models, based on statistical metrics, was conducted using a dataset featuring 7 key parameters, showcasing their accuracy in predicting WQI and classifying water quality effectively. Results indicated that NARNET slightly outperformed LSTM in WQI prediction, while SVM achieved the highest WQC prediction accuracy with 97.01%. During testing, LSTM and NARNET showed close accuracies with slight differences in regression coefficients with 96.17% and 94.21% respectively.

Finally, in a study conducted by [13] ML models were trained using the Water Quality dataset sourced from the Indian Government website via Kaggle. The WQI served as the basis for data categorization. Various ML algorithms, including DTs, MLP, XGBoost, KNN, and SVM, were investigated. To evaluate model effectiveness, metrics including Precision, Recall, Accuracy, and F1-Score were utilized. Results indicated that XGBoost exhibited superior performance as a water quality classifier, boasting an accuracy of 95.12%, closely followed by SVM with an accuracy of 93.22%.

While DL and ML approaches have proven effective in water quality assessment, several limitations exist in prior work. AutoDL, despite its automation, struggles with domain-specific optimizations, as conventional models outperform it by 1.8% in binary classification and 1% in multiclass classification. Additionally, model comparisons often lack robust justification, as seen in Barzegar et al., where SVR, DT, LSTM, CNN, and CNN-LSTM were analyzed for specific water quality parameters without evaluating fully connected architectures like MLP, which have shown strong performance in another research.

Furthermore, hybrid models like CLT introduce additional computational complexity, requiring feature selection via the GJO algorithm, making them less practical for real-time water monitoring. Similarly, supervised learning approaches using SMOTE, as in Dritsas and Trigka, achieve high accuracy (98.1%) but risk bias due to synthetic data introduction. Prior research highlights MLP's strength in regression tasks, with Najah et al. reporting an R^2 of 0.93, outperforming other models, while Wu and Wang demonstrated MLP's effectiveness in handling missing data alongside LSTM's ability to process sequential dependencies. Moreover, Jiang et al. showed that GRU outperformed LSTM and traditional ML models by 9.13%-15.03% in R^2, reinforcing the effectiveness of recurrent models for time series water quality prediction. These findings emphasize the need for a more balanced evaluation of model architectures, computational feasibility, and data augmentation risks in future research.

3. Research Methodology and approach

3.1. Background of the Research Study

Google collab was the platform for conducting this study; moreover, ML libraries in python called keras and Scikit-learn were used in the programming phase. Moreover, two ML techniques being LSTM, and MLP

were used on the dataset. This study followed a six-phase methodology, which is CRISP-DM [14].

Figure 1. Phases of the CRISP-DM Methodology.

3.2. *Dataset Description*

The dataset focuses on water quality (WQ) parameters in India, collected between 2012 and 2021. Data was gathered from an authorized Indian government website, comprising 7339 entries with six attributes per entry and a solitary outcome variable [15]. A breakdown of these attributes is detailed in Table 1.

When evaluating water drinkability various essential factors were considered, such as Biochemical Oxygen Demand (BOD), electrical conductivity (EC), pH, DO, and total coliforms (TC). A WQI is calculated using these parameters to provide a comprehensive evaluation of water potability. The WQI calculation involves deriving new parameters from the original measurements using a specific classification system. These derived parameters are denoted as ndo, nco, nbdo, nec, npH, and nna, representing normalized values for pH, DO, TC, BOD, EC, and NA. Weighted averages are then calculated for each parameter using the following formulas shown in equation (1-7) [15]:

- Weighted contribution for pH is given by:

$$wph = nph * 0.165 \quad (1)$$

- Weighted contribution for DO is given by:

$$wdo = npo * 0.281 \quad (2)$$

- Weighted contribution for Biochemical Oxygen Demand (BOD) is given by:

$$wbdo = nbdo * 0.234 \quad (3)$$

- Weighted contribution for Electrical Conductivity (CE) is given by:

$$wec = nec * 0.009 \quad (4)$$

- Weighted contribution for Sodium (NA) is given by:

$$wna = nna * 0.028 \quad (5)$$

- Weighted contribution for Total Coliform (TC) is given by:

$$wco = nco * 0.281 \quad (6)$$

- The Final Water Quality Index (WQI) is

Calculated using the formula:

$$\text{WQI} = wph + wdo + wbdo + wec + wna + wco \quad (7)$$

Water samples are categorized as drinkable (1) when the WQI equals or exceeds 75, and undrinkable (0) if the WQI falls below 75. This method of classification offers a uniform method for evaluating water suitability using defined benchmarks and measured concentrations.

Table 1. Dataset Description

Attribute	Description	Datatypes
Dissolved Oxygen (DO)	Optimum DO Concentration is 10 mg/L.	float64
pH	The required pH is 8.5.	float64
Conductivity (EC)	The wanted Conductivity is 1,000 µS/cm.	float64
Biological Oxygen Demand (BOD)	The optimum concentration is 5 mg/L.	float64
Nitrate (NA)	The optimum concentration is 45 mg/L.	float64
Total coliform (TC)	The required value is 100 per 100 mL	float64
Potability	A rating of 1 indicated that the water is safe to drink, while a rating of 0 means it is not safe to drink.	object

3.3. *Dataset Preparation*

After examining the data, the next step is to prepare it for analysis and model building. This involves addressing missing data, transforming categorical variables into numerical values, and dividing the dataset into appropriate testing and training subsets.

3.3.1. *Missing Data*

In this study, two essential functions commonly used to check for missing or duplicated data in a dataset, isnull().sum() and duplicated().sum(), were applied to ensure data integrity. The isnull().sum() function identifies and counts missing values in each column, while duplicated().sum() detects redundant rows. After executing these functions, the results confirmed that there were no missing or duplicated data, affirming the dataset's completeness and consistency. This ensures higher data quality, ultimately enhancing the reliability and accuracy of the ML model.

3.3.2. *Balancing the Dataset*

The value counts () function was used to assess class balance, revealing 3,958 instances for class 0 and 3,381 instances for class 1. A dataset is typically considered imbalanced if one class significantly outweighs the other.

To quantify this, the imbalance ratio was calculated as 3958 / 3381 ≈ 1.17, indicating that the dataset is relatively balanced. Generally, datasets with imbalance ratios exceeding 1.5–2.0 are considered imbalanced and may require techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or undersampling to correct class distribution.

3.3.3. *Encoding Categorical Data*

The LabelEncoder function was applied to the dataset to convert categorical data into numerical values, a crucial step for preparing data for ML models, which typically require numerical inputs. In this study, water samples were classified based on the WQI, where a value of 75 or higher indicated potable water (labeled as 1), while a WQI below 75 signified non-potable water (labeled as 0). This transformation ensures that the dataset is properly formatted for model training therefore enhancing the efficiency of the classification process [16].

3.3.4. *Splitting Data*

To assess the capabilities of the algorithms at water potability forecasting, the study employed a 10-fold cross-validation approach. This method ensured a robust *assessment of the algorithms, improving the generalizability and reliability of the research discoveries [17].*

3.3.5. *Data Normalization*

The numerical data was normalized to scale values within a predefined range, typically 0 to 1 or -1 to 1, ensuring uniform feature contribution. This transformation prevents any single variable from dominating the model, promoting balanced and unbiased learning.

3.4. *Modelling*

The two NNs, MLP and LSTM are implemented to predict water potability.

MLPs: a type of ANN, excel at solving complex classification and regression tasks. These networks are structured as layers of interconnected nodes, each processing information and relaying it to the next layer. The network learns by fine-tuning the connections between nodes, enabling it to recognize complex patterns within data. MLPs are particularly adept at handling non-linear relationships and can be trained to achieve high accuracy across a wide range of applications [18].

In this study The MLP model had been configured with 2-3 hidden layers, with each layer having 64, 128, or 256 neurons, depending on the complexity of the task. The activation functions used had been ReLU for the hidden layers to facilitate efficient training and Sigmoid for the output layer to support binary classification. The Adam optimizer had been selected for its adaptive learning properties, while the Binary Crossentropy loss function had been utilized to optimize performance for the binary classification task. The batch size had been set to 64 to balance computational efficiency and memory usage. The model had been trained over 50 to 100 epochs, with a learning rate of 0.001 to ensure stable convergence. Additionally, dropout rates between 0.2 and 0.3 had been applied to prevent overfitting during training.

LSTM: a type of RNN, designed to handle sequential data. Unlike traditional RNNs, it utilizes a unique memory cell structure that allows them to retain information over extended periods. This renders them well-suited for assignments encompassing natural language processing, time series analysis, and speech recognition. LSTMs' ability to "remember" past information enables them to grasp extended dependencies within sequences, leading to improved accuracy in predicting future outcomes [19].

In this study the LSTM model had been configured using a network architecture comprising 1-2 LSTM layers followed by a Dense layer. The number of units (neurons) in each LSTM layer had ranged from 64 to 256, depending on the data's complexity. The activation functions used had been Tanh for the LSTM layers and Sigmoid for the output layer, facilitating efficient learning and binary classification. The Adam optimizer had been employed for its adaptive learning capabilities, paired with the Binary Cross entropy loss function for optimizing binary classification tasks. The batch size had been set to 64, balancing memory usage and computational efficiency. The model had been trained over 50 to 100 epochs, with dropout rates between 0.2 and 0.5, and recurrent dropout rates ranging from 0.2 to 0.3 to reduce overfitting. Finally, the learning rate had been set to 0.001 to ensure steady convergence during training.

3.5. *Performance Evaluation*

The capabilities of the NNs are identified through accuracy, sensitivity, precision, and ROC-AUC are all used to measure their performance.

3.5.1. *Accuracy:*

This metric represents the portion of true forecast by a model out of all predictions made as shown in equation (8) [20].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

3.5.2. *F-measure:*

The F-measure provides an equitable evaluation of a model's precision and recall. It calculates a weighted average of these two metrics, providing an extensive assessment of the model's capabilities shown in equation (9) [21].

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (9)$$

3.5.3. ROC-AUC Value:

It evaluates a classification model's performance across various threshold settings. The AUC indicates the model's ability to distinguish between classes, and higher ROC-AUC scores signify improved classification performance and discriminative capacity [21].

3.5.4. Precision:

It assesses the ratio of accurately identified negative instances (true negatives) among all cases predicted as negative shown in equation (10) [21].

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

3.5.5. Recall:

It evaluates a model's capacity to accurately detect all positive instances present in a dataset shown in equation (11) [21].

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

4. Results

The experimental findings revealed that MLP NN model outperformed the LSTM NN model in WQC tasks. The MLP model achieved a near-perfect accuracy of 99.9%, significantly higher than the 97.6% accuracy of the LSTM model. This suggests that the MLP is more reliable in correctly classifying the water potability samples.

The MLP's F-measure of 99.9% surpasses the LSTM's 97.1%. Recall and Precision for the MLP model are near perfect at 99.9%, indicating that the MLP has a very low false positive and false negative rate. The LSTM, while still performing well, shows slightly lower precision (97.1%) and recall (97.5%), suggesting it is less effective in minimizing these errors as shown in Table II and Figure 2.

This ROC-AUC curve comparison highlights the classification performance of the MLP and LSTM models. The MLP model achieves a perfect AUC of 100%, represented by a diagonal line, indicating flawless classification, whereas the LSTM model attains an AUC of 97.9%, demonstrating strong but slightly lower performance as shown in Figure 3.

Table 2. Performance Comparison between NNs.

Models	MLP	LSTM
Accuracy	99.9%	97.6%
F-measure	99.9%	97.1%
Precision	99.9%	97.1%
Recall	99.9%	97.5%
ROC-AUC	100%	97.9%

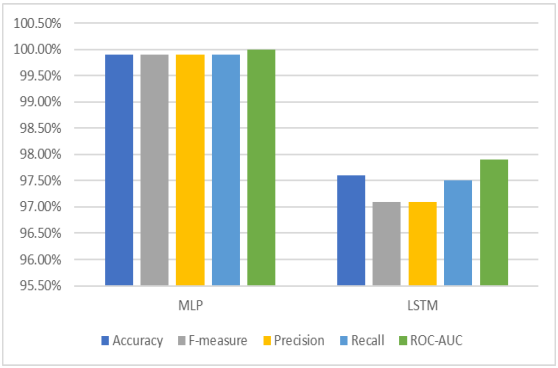

Figure.2: Performance Plot of Proposed Neural Networks

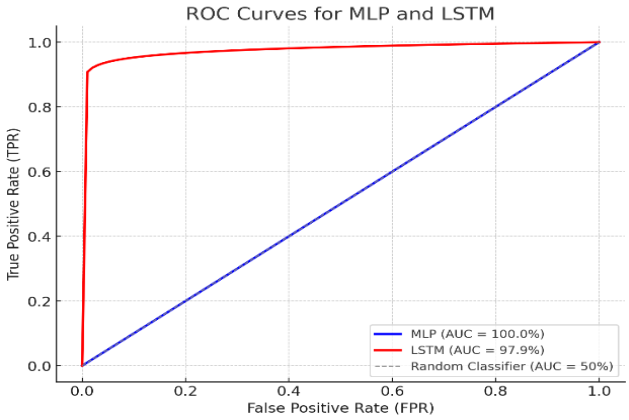

Figure.3: The ROC-AUC Plot of Proposed Neural Networks

5. Discussion

The current study demonstrates the superiority of MLP over LSTM for binary water potability classification, with MLP achieving 99.9% across all metrics, outperforming results from prior studies. Compared to Prasad et al., who found conventional DL models achieving up to 99% accuracy, and Dritsas and Trigka, whose stacking model reached 98.1%, the MLP model's perfect ROC-AUC of 100% sets a new benchmark. While LSTM performed slightly below MLP in this study, previous research, such as by Barzegar et al. and Liu et al., highlights LSTM's strength in multivariate and temporal tasks, particularly in hybrid architectures like CNN-LSTM and CLT. Studies like Najah et al. and Wu and Wang reaffirm MLP's excellence for regression and classification tasks, and the current study underscores its efficiency for simpler tasks without relying on hybrid approaches or extensive preprocessing. This suggests that while LSTM excels in capturing temporal dependencies, MLP's straightforward architecture is more effective for binary classification.

Many previous studies have focused on multi-class classification or regression-based approaches to predict the WQI rather than directly classifying water as potable or non-potable. For instance, Liu et al. employed a hybrid CLT model to predict multivariate water quality parameters, concentrating on continuous values instead of binary classification. Similarly, Barzegar et al. used a CNN-LSTM hybrid model to predict DO and Chl-a levels rather than focusing on binary classification. These models require additional steps to convert their outputs into potable/non-potable labels, adding unnecessary complexity and potential misclassification risks. Consequently, there is a need for direct binary classification models that efficiently determine water potability. The current research explicitly focuses on binary classification, optimizing the process for direct decision-making and achieving a near-perfect accuracy of 99.9% using the ML model.

Many prior studies have also overcomplicated their models by integrating multiple DL architectures (e.g., CNN-LSTM, ANN-WT-LSTM) under the assumption that this will enhance predictive accuracy. Wu and Wang, for example, used a hybrid ANN-Wavelet Transform-LSTM model, significantly increasing computational complexity. Similarly, Najah et al. combined IoT with ML models such as LSTM, SVR, MLP, and NARNet for WQI prediction, instead of opting for a simpler and more effective classifier. These hybrid models tend to be computationally expensive, require extensive hyperparameter tuning, and lack interpretability, making them impractical for real-time applications. Additionally, many studies fail to demonstrate whether the added complexity truly results in better performance than simpler models like MLP. In contrast, the current research demonstrates that a straightforward MLP model can outperform LSTM, achieving an accuracy of 99.9% compared to LSTM's 97.6%, without requiring hybrid architectures.

Another major limitation in prior work is the neglect of feature selection and data preprocessing. Some studies feed raw data directly into DL models without conducting proper feature selection or normalization. For example, Liu et al. applied GJO for feature selection, but many other studies lacked systematic feature engineering. The absence of feature selection can lead to overfitting and reduced generalizability. Moreover, many prior works do not systematically explore how feature normalization or selection impacts model performance. The current study addresses this gap by applying proper feature engineering, ensuring that only the most relevant attributes contribute to model performance, thereby improving accuracy and efficiency.

Additionally, many studies have overlooked traditional ML models, assuming that DL models always perform better. For instance, Aldhyani et al. compared SVM, KNN, and Naïve Bayes with DL models, but many other studies did not conduct such benchmarking. Traditional ML models, such as DT, RF, and SVM, can sometimes perform equally well or even better than DL models, particularly with smaller datasets. Many studies fail to justify why DL is necessary over simpler, more explainable models. In contrast, the current study provides a clear justification for using DL.

The superior performance of MLP over LSTM in this study can be attributed to the nature of the task and the architectural differences between the two models. MLP, being a feedforward NN, is well-suited for binary classification tasks where the data lacks significant temporal dependencies. Its simpler architecture focuses on mapping inputs directly to outputs through fully connected layers, enabling efficient learning of non-linear relationships in static datasets.

By critically analyzing prior work, the selection of MLP and LSTM had been justified as offering high predictive accuracy, robust performance in both regression and classification tasks, and efficient processing for multivariate tabular data and time-series forecasting. Additionally, better scalability for real-world applications had been provided compared to computationally expensive hybrid models. The ability to handle both static and sequential data effectively while outperforming existing DL models in accuracy and efficiency had made them optimal choices for water quality assessment.

6. Conclusion a Future Direction

The results clearly demonstrate the inferior performance of the LSTM model contrasted to the superior MLP model in water potability classification. Across all metrics, the MLP consistently outperforms the LSTM, achieving near-perfect scores for accuracy, F-measure, precision, recall, and a perfect ROC-AUC score. This suggests that the MLP's architecture is more fitted for capturing the intricate relationships and patterns present in the datasets concerning water potability, leading to more precise predictions.

This study highlights the crucial role of NNs in classification processes, particularly for complex datasets like those found in water potability analysis. NNs, utilizing their capability to comprehend intricate patterns and adjust to non-linear connections, offer a powerful tool for tackling such challenges. The MLP's superior performance in this instance underscores the importance of selecting the right NN architecture for the specific task since each model excel in different domains.

Future research can advance water potability classification by integrating hybrid DL models, such as

combining MLP with CNN or Transformer-based architecture, to enhance feature extraction and classification accuracy. Utilizing advanced feature selection techniques, including genetic algorithms and particle swarm optimization, can further refine model performance. Expanding from binary to multi-class classification would allow for a more detailed evaluation of water quality levels. Implementing IoT-enabled real-time monitoring systems can enable continuous water quality tracking and instant alerts when contamination exceeds safety thresholds. Additionally, incorporating spatiotemporal analysis using GIS and remote sensing data can improve predictive capabilities. Exploring diverse datasets from various geographical regions and environmental conditions would enhance model robustness, ensuring its applicability across different water sources and pollution levels.

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] J. J. Bogardi, J. Leentvaar, and Z. Sebesvári, "Biologia Futura: integrating freshwater ecosystem health in water resources management," *Biologia futura*, vol. 71, no. 4, pp. 337-358, 2020, DOI: doi.org/10.1007/s42977-020-00031-7.

[2] N. I. Obialor, "The Analysis of Water Pollution Control Legislation and Regulations in Nigeria: Why Strict Implementation and Enforcement Have Remained a Mirage," *Available at SSRN 4600394*, 2023.

[3] R. Huang, C. Ma, J. Ma, X. Huangfu, and Q. He, "Machine learning in natural and engineered water systems," *Water Research*, vol. 205, p. 117666, 2021, DOI: 10.1016/j.watres.2021.117666.

[4] D. V. V. Prasad *et al.*, "Analysis and prediction of water quality using deep learning and auto deep learning techniques," *Science of the Total Environment*, vol. 821, p. 153311, 2022, DOI: 10.1016/j.scitotenv.2022.153311.

[5] R. Barzegar, M. T. Aalami, and J. Adamowski, "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model," *Stochastic Environmental Research Risk Assessment*. vol. 34, no. 2, pp. 415-433, 2020, doi.org/10.1007/s00477-020-01776-2.

[6] W. Liu, T. Liu, Z. Liu, H. Luo, and H. Pei, "A novel deep learning ensemble model based on two-stage feature selection and intelligent optimization for water quality prediction," *Environmental Research*, vol. 224, p. 115560, 2023, DOI: 10.1016/j.envres.2023.115560.

[7] 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.

[8] A. Najah, A. El-Shafie, O. A. Karim, and A. H. El-Shafie, "Application of artificial neural networks for water quality prediction," *Neural Computing Applications*,vol. 22, pp. 187-201, 2013, DOI: 10.1007/s00521-012-0940-3.

[9] J. Wu and Z. Wang, "A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory," *Water*, vol. 14, no. 4, p. 610, 2022, DOI: 10.3390/w14040610.

[10] Y. Jiang, C. Li, L. Sun, D. Guo, Y. Zhang, and W. Wang, "A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks," *Journal of Cleaner Production*, vol. 318, p. 128533, 2021, DOI: 10.1016/j.jclepro.2021.128533.

[11] M. Hmoud Al-Adhaileh and F. Waselallah Alsaade, "Modelling and prediction of water quality by using artificial intelligence," *Sustainability*, vol. 13, no. 8, p. 4259, 2021, DOI: 10.3390/su13084259.

[12] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, " Water Quality Prediction Using Artificial Intelligence Algorithms," *Applied Bionics Biomechanics*, vol. 2020, no. 1, p. 6659314, 2020, DOI: 10.1155/2020/6659314.

[13] J. Kirui, "Machine Learning Models for Drinking Water Quality Classification," in *2024 International Conference on Control, Automation and Diagnosis (ICCAD)*, 2024, pp. 1-5: IEEE, DOI: 10.1109/ICCAD60883.2024.10553712.

[14] 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.

[15] 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/150184.

[16] 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.

[17] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction," *Cleaner Engineering Technology*,vol. 15, p. 100664, 2023, DOI: 10.1016/j.clet.2023.100664.

[18] S. Afzal, B. M. Ziapour, A. Shokri, H. Shakibi, and B. Sobhani, "Building energy consumption prediction using multilayer perceptron neural network-assisted models; comparison of different optimization algorithms," *Energy*, vol. 282, p. 128446, 2023, DOI: 10.1016/j.energy.2023.128446.

[19] K. Li, W. Huang, G. Hu, and J. Li, "Ultra-short term power load forecasting based on CEEMDAN-SE and LSTM neural network," *Energy Buildings*,vol. 279, p. 112666, 2023, DOI: 10.1016/j.enbuild.2022.112666.

[20] 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.

[21] F. A. Musleh and R. G. Taha, "Forecasting of forest fires using machine learning techniques: a comparative study," 2022, DOI: 10.1049/icp.2023.0571.

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 and the Second Prize for Best Use of Satellite Data in the competition held by the NSSA in celebration of the launch of Bahrain's first satellite, Al Munther.

Dr. Fuad Musleh is an Assistant Professor at the University of Bahrain. He received his Ph.D. and MSc degrees from the University of Alabama in Huntsville, B.Sc. from Jordan University of Science and Technology in Jordan. He is interested in research related to flow through vegetation, water and environmental conservation. Additionally, he is interested in leveraging Data Analysis in his field.

Mr. Abdel Rahman Musleh is a Senior BSc in Electrical Engineering student at the University of Bahrain, with an interest in leveraging AI and ML to develop Cyber-Physical Systems that enhance efficiency, reliability, and sustainability.

Received: 04 April, 2025, Revised: 07 May, 2025, Accepted: 15 May, 2025, Online: 25 May, 2025

DOI: https://doi.org/10.55708/js0405002

Cavity Sensing for Defect Prevention in Injection Molding

Oumayma Haberchad [*1], **Yassine Salih-Alj** [2]

[1] Control and Instrumentation Engineering Department, College of Engineering and Physics, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia

[2] School of Science and Engineering, Al Akhawayn University, Ifrane, 53000, Morocco

Email: Oumayma Haberchad (o.haberchad@aui.ma) , Yassine Salih-Alj (y.salihalj@aui.ma)

*Corresponding author: Oumayma Haberchad, Dhahran, Saudi Arabia, +966537242208 & g202423940@kfupm.edu.sa

ABSTRACT: Real-time monitoring of injection molding parameters plays a pivotal role in enhancing product quality, reducing defects and improving production. This study presents a cavity data acquisition system for real time monitoring of process parameters inside the mold. The system consists of non-destructive in-mold sensors that monitor the status of the melt within the cavities. Furthermore, the geometry of the injected part is taken into consideration when selecting the position of the sensors. This enables early discovery of defects by studying abnormal variations of the monitored parameters in areas where these defects are suspected. A case scenario is shown in which we simulate the molding profile of a plastic part using SolidWorks Plastics. The suggested sensors' placements are then derived. Results indicate that the piezoelectric sensor measures with a root mean square error (RMSE) that is less than 0.0004 V and a peak error of 0.0012 V. The proposed method promises more control over injection conditions inside the mold, as well as enhanced overall production quality.

KEYWORDS: Injection Molding, Sensors, Data Acquisition, Process Monitoring; Quality Control.

1. Introduction

Since the invention of polymers, plastic products have been widely employed to meet basic demands and to replace expensive materials while providing competitive performance. Plastic's high popularity stems from its adaptability and capacity to mimic the functions of other materials while offering enhanced features such as corrosion resistance, low weight, flexibility, inexpensive production and maintenance costs [1].

Plastic manufacturing has significantly been used in many industries. The use of plastic in the automobile and aircraft industries led to a reduction in fuel consumption thanks to its light weight. Its usage in the medical industry improved safety and reduced contamination caused by metal equipment [2].

Different processes are used to produce plastic parts. This includes extrusion, thermoforming, blow molding, and injection molding.

Injection molding is a popular manufacturing technology. It is widely used in various industries, including automotive, medical, and electronics. The injection molding market have reached 305 billion USD in 2022 only, accounting for over one-third of total plastic manufacturing [3]. The sector has evolved significantly in recent decades. Concerns over plastic waste and Industry 4.0's demand for digitalized processes, including cyber-physical systems, are driving these changes [4]. To address these challenges, different methods have been proposed to automate and optimize the injection molding process [5], [6].

To fully regulate the injection molding process, machine and in-mold parameters, as well as part quality, must be monitored [7]. Tracking machine parameters offers real-time insight into the polymer's behavior during dehumidification, melting, and injection phases. Previous works have analyzed data from the injection machine to determine which parameters have the greatest impact on the quality of ejected parts [8]. This strategy produced notable outcomes. However, injection molding's non-iterative nature led to varying results across multiple manufacturing cycles.

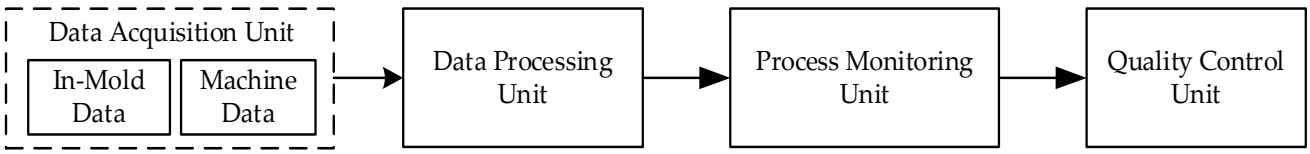

Figure 1: Block diagram of the considered cavity data acquisition system.

In mold process monitoring entails tracking the polymer melt (resin) during injection, cooling, and ejection. Previous works in the literature suggest that extracting data from a mold can provide better insights about melt status and behavior [7], [9].

Melt enters mold cavities later in the molding cycle. This stage involves sophisticated procedures including cooling, ejection, and most importantly, part formation. The polymer's condition can be closely monitored as it transitions from liquid to solid using in-mold sensors. While sensors have been used in machines since the 1960s, cavity sensors represent a recent technological adoption (10-12 years) [10]. This is due to the complex mounting of these sensors inside the mold which delays their mass integration.

However, with an increasing migration towards automated monitoring of industrial processes, cavity sensing technology has gained more popularity, and an array of new sensors has been developed to ensure agile integration into machinery.

Cavity sensors typically use piezoelectric and strain gauges for pressure measurement and thermocouples to monitor temperature [11]. These are available in various sizes and mounting options, allowing for full customization to meet quality standards. However, numerous limitations have been affecting their performance. For instance, sensors mounted in molds can be damaged by high pressure and temperature changes, leading to corrosion and frozen layers. This results in measuring inaccuracies. Additionally, while direct contact with the melt can result in accurate measurements, it may also cause surface level defects in the final product.

Thin-film sensors have also been investigated. They are made of piezoelectric sensors that are deposited into the steel surface via sputtering, allowing for precise pressure monitoring in various parts of the mold [7]. Temperature changes, however, significantly disrupt the signals affecting measurement accuracy.

Other alternative sensors have been proposed, including wireless piezoelectric sensors [12], infrared sensors [13], and optical sensors [14]. Although they are still in their early stages of implementation in industry, they provide efficient and precise measurements without direct contact with the melt which expands the range of parameters that can be monitored.

To allow full integration of cavity sensors with the other components of the mold, different Data acquisition systems have been developed. This includes an Arduino microcontroller-based data acquisition module that allows the visualization of different mold parameters including mold temperature, cavity pressure, 3-axis acceleration, and extraction force [15]. These were measured based on commercially available sensors including thermocouples, pressure sensors, and force sensors. The developed system allowed affordable, simple, and real-time data acquisition and monitoring of process parameters. Additionally, the system was able to distinguish between normal and abnormal patterns in monitored parameters. Although process variations were successfully captured, the wired nature of sensors used can cause potential hazards due to mold's movements. To overcome this issue, a multiple measurement sensor was adopted in [16] to measure temperature and pressure simultaneously. The sensor was equipped with a piezoelectric transducer for pressure measurement, and a K thermocouple for temperature measurement. This reduced the amount of holes required to insert sensors inside the cavities thereby minimizing structural damage to the mold. Other commercially available controllers include Priamus' Fillcontrol, BlueLine Hardware and QFlow Systems Engineering.

Considering cavity sensors' high efficiency in monitoring injection molding processes and addressing the lack of sufficient work recognizing the geometry of the part when selecting sensors, this work presents a data acquisition system for in-mold process control and monitoring. The system allows the measurement of injection molding process parameters inside the mold using cavity sensors. The main objective of the study is to select sensors capable of accurately monitoring cavity process parameters in areas where defects are suspected while taking into account the geometry of the part and without inducing destructive alterations to the mold. To achieve this, ultrasonic based sensors were proposed. These sensors are positioned based on the part's molding profile to identify abnormal melt flow and reduce defects. Thank to their wireless transmission scheme, these sensor reduce the amount of wholes and wires inserted into the mold resulting in a more robust process monitoring. The proposed approach can promote sustainable injection molding by decreasing waste and adjusting to future changes in the manufacturing process.

The remainder of the paper is structured as follows. Section 2 describes the theoretical model for cavity data acquisition. Section 3 presents the proposed data acquisition system.

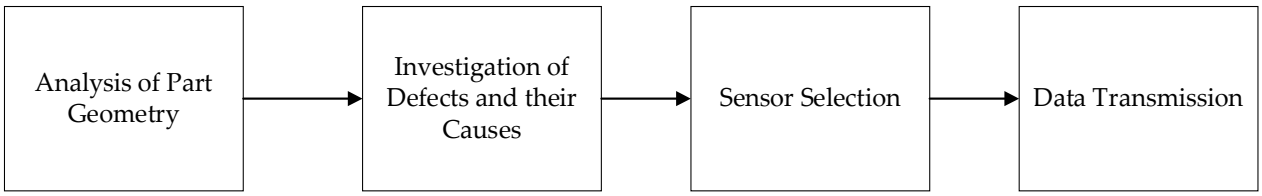

Figure 2 : System model of the data acquisition system.

Section 4 details the obtained simulation results. Finally, Section 5 concludes the paper.

2. Theoretical Model for Cavity Data Acquisition

This section presents the injection molding process cycle data flow. Figure 1 depicts four levels of data processing: data acquisition unit, data processing unit, process monitoring unit, and quality control unit [17].

2.1. Data Acquisition Unit

Data collection begins with measuring process parameters using sensors incorporated in both the injection machine and the mold. The machine is equipped with a complex set of sensors that allow for full monitoring of various process parameters such as maximum injection pressure, screw position, and temperature in different screw zones.

Due to direct contact with the melt, in-mold sensors are incredibly effective in describing melt flow behavior while being less complex. For instance, these sensors can measure mold temperature and pressure, melt flow, velocity, and viscosity [7].

2.2. Data Processing Unit

The data obtained from injection molding machines and in-mold sensors is then processed. The collected signals are amplified and filtered to remove noise and enhance the data. The continuous signals generated by the sensors are then sampled to improve computational efficiency and synchronize the numerous data sources.

2.3. Process Monitoring Unit

At this stage, the processed datasets are utilized to monitor injection cycles. Machine and mold parameters are visualized via graphical interfaces, allowing operators to track variations in process parameters as well as melt rheology to ensure that no disruptions influence production [6].

2.4. Quality Control

Once the part is ejected out of the mold, it undergoes multiple control procedures. First, when the injection cycle ends, and later once the visual and dimensional features of the part stabilize. This ensures that the part meets the standards set by the customer.

3. Proposed Data Acquisition System

In this section a detailed description of the proposed data acquisition system is given. As illustrated in Figure 2, four steps are to be implemented to develop the model: analysis of part geometry, investigation of defects and their causes, sensor selection, and data transmission.

3.1. Analysis of Part Geometry using CAD Software

Part design is the first step in the mold creation process. Computer-aided design simulations, such as Autodesk, Ansys, and SolidWorks, have made part design more automated. Injection molding software can design parts and simulate melt behavior. This provides valuable insights into how parameters vary during the filling process. Additionally, it allows for early detection of defects during the injection process.

3.2. Analysis of Defects and Their Causes

After the part is manufactured, the next step in implementing the suggested cavity data acquisition system is to analyze the injection molding defects associated with the mold. Weld lines, shrinkage, warpage, and sink marks are some of the most prevalent defects.

3.2.1. Weld Lines

Weld lines are plastic flow traces that resemble the letters J or U. This defect occurs when two fronts flow from different directions meet, resulting in weak regions in the component [18].

3.2.2. Shrinkage

Shrinkage is the reduction in volume caused by polymer cooling. Inconsistent contraction due to temperature variations in the various regions of the part causes shrinkage [19]. Despite its common occurrence, excessive shrinkage can cause geometric errors in plastic parts.

The shrinkage can be expressed using equation (1), [20]:

$$s_i = \frac{d_i - d_{si}}{d_i}, \tag{1}$$

where s is the shrinkage, d is the cavity width, and d_s is the part's width. The i subscript indicates the sensor's position. Using different sensors, temperature and pressure can be measured locally to calculate the overall shrinkage of the part.

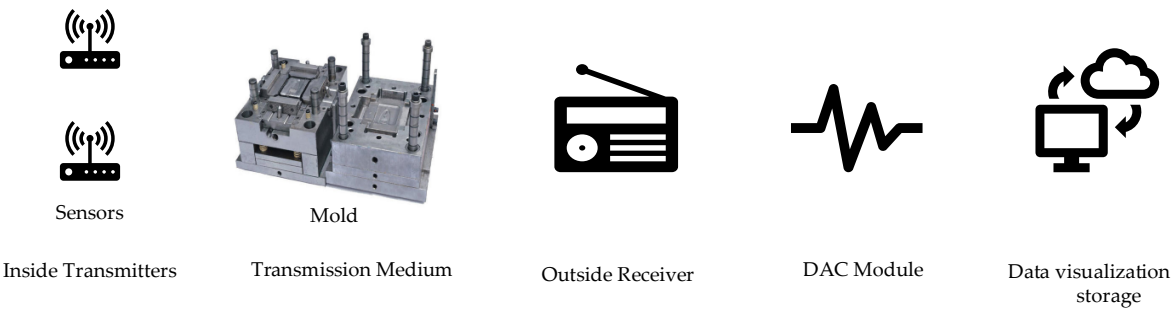

Figure 3 : Transmission scheme of data extracted from cavity sensors.

3.2.3. *Warpage*

Warpage occurs when internal forces cause a part to bend and deviate from its original geometry. Warpage, as is shrinkage, is induced by uneven part contraction and ongoing cooling after ejection, which causes parts with slower cooling rates to bend [19].

3.2.4. *Sink Marks*

Sink marks appear on surfaces with substantial wall thickness on the opposite side [18]. Such defects could occur due to fluctuations in cavity temperature, in addition to variations in other process factors including cooling time and packing pressure [21].

3.3. *Sensor Selection*

Mold sensors monitor numerous parameters inside the cavity, including pressure and temperature. This work considers sensors that take into account both defects and part geometry. To avoid damaging the mold during installation, two sensors are proposed: a wireless piezoelectric sensor and an ultrasonic sensor.

3.3.1. *Wireless Piezoelectric Sensors*

These sensors monitor pressure by transmitting energy from melt pressure, which also powers them [12]. An energy converter converts mechanical energy from the melt into electrical energy. The energy converted is expressed as follows:

$$E = \frac{1}{2} C \left(\frac{n \cdot d_{33} \cdot P \cdot A_{fp}}{C} \right)^2, \qquad (2)$$

where C is the total capacitance, d_{33} is the charge constant, P is the pressure acting on the piezoceramic ring, A_{fp} is the footprint area where the pressure is acting, and n is the number of piezoceramic rings. For instance, electric charge is generated under mechanical stress induced by melt pressure. The resulting charge can be expressed as the product of the pressure acting on the ring, the footprint area over which the pressure is acting, and the charge constant. The electrical behavior of the piezoelectric rings can be approximated as a parallel plate capacitor. The ratio of the charge over the total capacitance of the system models the voltage generated, and it is related to the square root of the energy as expressed in equation 2.

The resulting electrical energy is discretized into electrical pulses using a threshold modulator. A signal transmitter converts the pulses into ultrasonic waves, which are subsequently delivered to a receiving unit outside the cavity.

3.3.2. *Ultrasonic Sensors*

Ultrasonic sensors are non-destructive and can measure parameters including melt homogeneity, temperature, and thickness [7]. Ultrasonic transducers use the converse piezoelectric effect to propagate ultrasonic waves [22]. Equations (2), (3), and (4) indicate how longitudinal ultrasonic velocity can be related to pressure and temperature using specific volume [23].

$$c_L = \left(\frac{1}{\rho \kappa} \right)^{\frac{1}{2}}, \qquad (3)$$

where c_L is the longitudinal ultrasonic velocity, ρ is the density of the polymer melt, and κ is the adiabatic compressibility expressed as:

$$\kappa = -\frac{1}{v} \left[\left(\frac{\partial v}{\partial P} \right)_T + \frac{T}{c_P} \left(\frac{\partial v}{\partial P} \right)_P^2 \right], \qquad (4)$$

where P is the melt pressure, T is the melt temperature, cp is the specific heat capacity, and v is the specific volume described by the Tait equation as [23]:

$$v(\mathrm{T}, \mathrm{P}) = v_0(T) \left[1 - C \cdot ln \left(1 + \frac{P}{B(T)} \right) \right] + v_t(T, P), \qquad (5)$$

where v_0 is the zero pressure isotherms, $B(T)$ is a temperature dependent function, and C= 0.0894 is a universal constant [24].

As the melt temperature increases, the sound velocity decreases in an approximately linear manner. On the other hand, increased melt pressure drives sound velocity to higher levels.

3.4. *Data Transmission*

Given the sensors used, the data acquisition scheme is based on the transmission of ultrasonic waves, allowing wireless communication between the transmitter and receiver [25]. To illustrate, when cavity parameters are

monitored, they are transformed into a voltage, which is then discretized into pulses using a threshold modulator [12]. The pulsing voltage causes the piezoelectric material to be displaced, resulting in ultrasonic waves [22]. These waves can travel through the mold's walls. Then they are received by an external receiver, which turns them back into voltage. The cavity measurements are recovered by multiplying the number of received ultrasonic pulses by the modulator's threshold. To display and save the measurements, a data acquisition module (DAC) can be employed to transform the analogue signal to a digital one. Figure 3 shows the overall transmission scheme.

4. Simulation Results

4.1. Simulation Setup

To visualize the proposed system's mechanisms, a computer-aided engineering (CAE) simulation has been constructed in SolidWorks Plastic. We began by designing the part, as shown in Figure 4. The part investigated was designed to cause the previously mentioned defects.

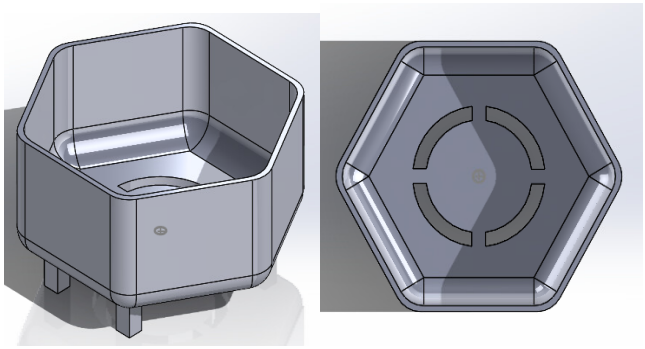

Figure 4 : CAD design of the studied part.

Then we proceed with the plastic flow analysis. The latter begins by defining the materials used. The chosen polymer is Acrylonitrile Butadiene Styrene (ABS), and its parameters are described in the SolidWorks plastic database, as seen in Table 1. This is followed by defining process parameters values for the simulation as expressed in Table 2. After that, the injection gate (through which the melt enters the cavity) was placed in the center of the part as seen in Figure 5.

Table 1: Material properties of ABS.

Property	Value
Melt Flow rate	35 g/10min
Max shear rate	50000 1/s
Max shear stress	0.3 MPa
Poisson's Ratio	0.39
Elastic Modulus	2250
Melt Temperature	230 °C
Max. Melt Temperature	280 °C
Min. Melt Temperature	200 °C
Mold Temperature	50 °C
Ejection Temperature	90 °C

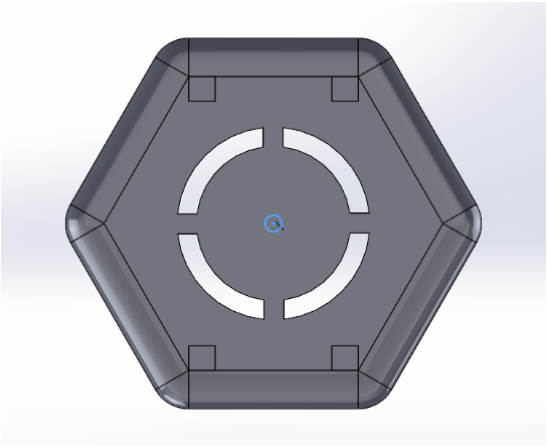

Figure 5 : Injection gate location.

Furthermore, a solid mesh with a total of 17410 triangular components measuring 5.19 mm was created as illustrated in Figure 6.

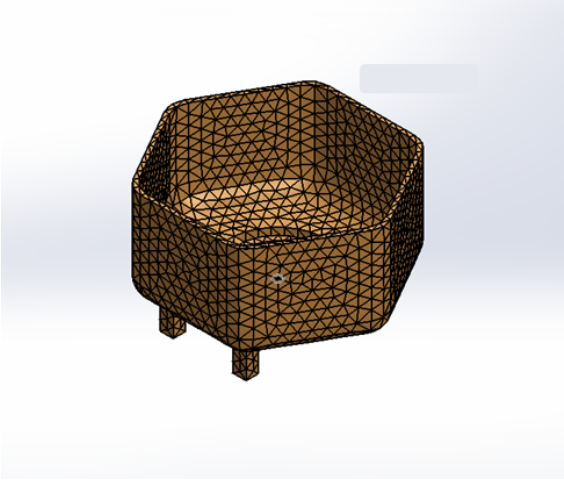

Figure 6 : Illustration of meshed part.

The simulation is then run, and the distribution of the various parameters and defects are shown.

Table 2: Simulation process parameter values.

Process parameter	Value
Melt temperature	230 °C
Mold temperature	50 °C
Injection pressure limit	100 MPa
Pure cooling time	31.163 s

4.2. Parameter Visualization

This section visualizes the various variables that control the process. The data was acquired after

performing the plastic flow simulation in SolidWorks Plastics.

In Figure 7, the max inlet pressure is visualized. We notice that the pressure increases until it reaches a maximum value of 13.488 MPa and then it decreases.

In Figure 8, the melt front flow rate versus time is visualized. We notice that the melted front flow rate varies increasingly until reaching a maximum value of 21.883 cc/s.

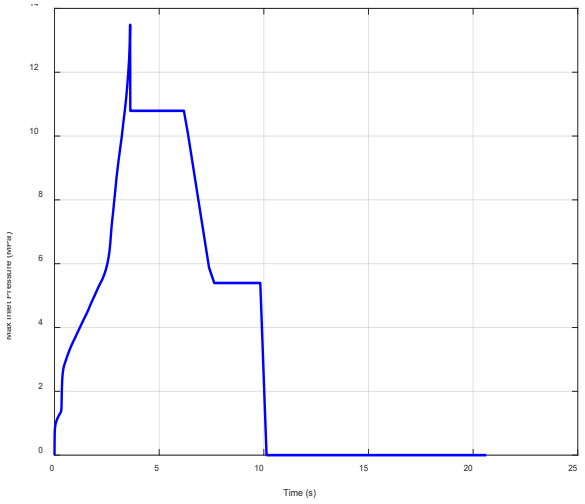

Figure 7 : Max inlet pressure versus time.

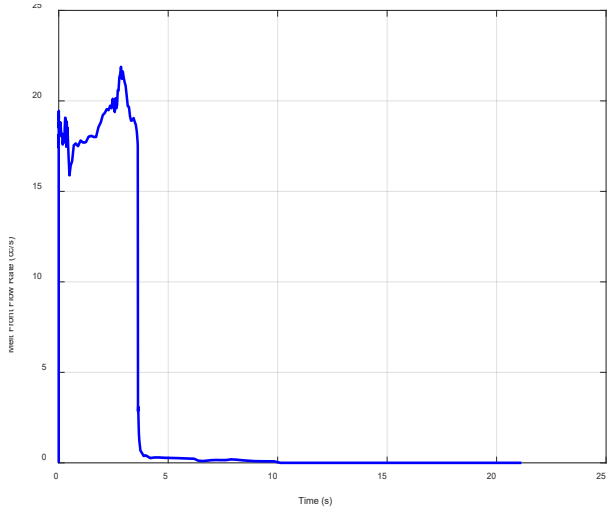

Figure 8 : Melt front flow rate versus time.

4.3. Analysis of Defects

In addition to analyzing the distribution of parameters within the cavity, fill and warp analysis enable us to identify potential defects that may appear during the molding process. The previously reported defects had been identified within the part after the simulations were completed.

Figure 9 depicts the distribution of sink marks in the portion. We notice that these marks are located in the lower area of the part because it has a high wall thickness, as well as in the four extruded pieces bearing the part, which have a high thickness and generate a depression on the other side of the part.

Figure 10 demonstrates the distribution of volumetric shrinkage. We observe that both the upper part and the gate endure significant shrinkage.

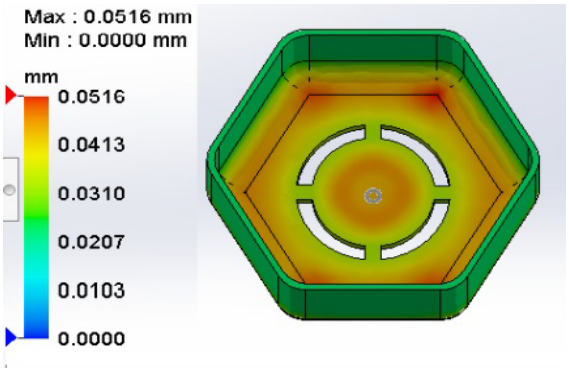

Figure 9 : Distribution of sink marks.

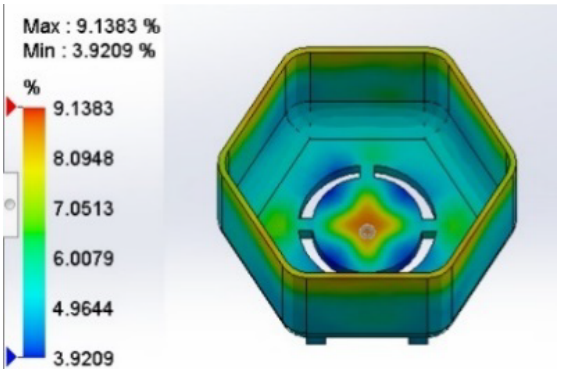

Figure 10 : Distribution of the volumetric shrinkage.

Figure 11 depicts the distribution of weld lines throughout the part. We notice that the weld lines are placed near the holes in the part, indicating that two melt flow fronts intersect in those areas.

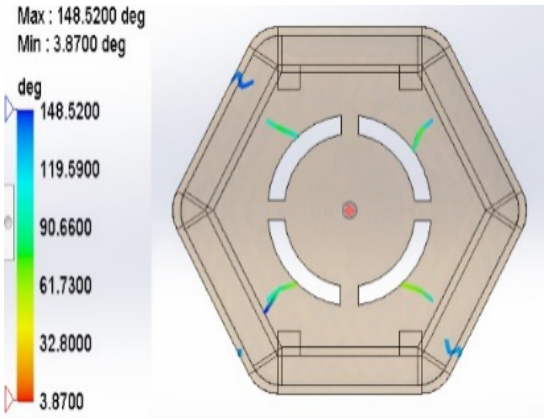

Figure 11 : Distribution of weld lines.

Figure 12 depicts the distribution of the total displacement of the part. The walls experience moderate warp levels, while the lower area experience greater warp.

4.4. Sensor's Placement

The system requires a total of five sensors, as shown in Figure 13. One wireless piezoelectric sensor will be installed where a sink mark is suspected in order to monitor pressure changes. Three more wireless piezoelectric sensors are utilized to monitor the part's shrinkage and displacement: one in the part wall, one at the top, and one at the bottom. One ultrasonic transducer

is used to measure temperature in the area where weld lines may appear.

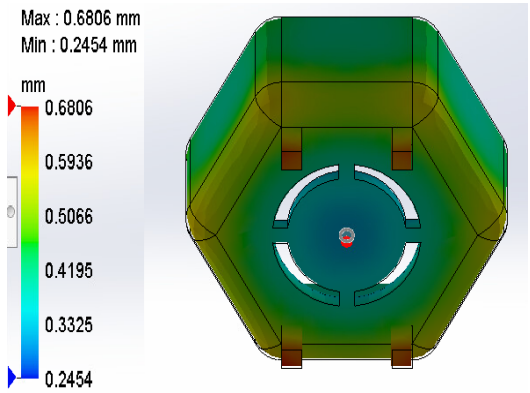

Figure 12 : Total stress displacement.

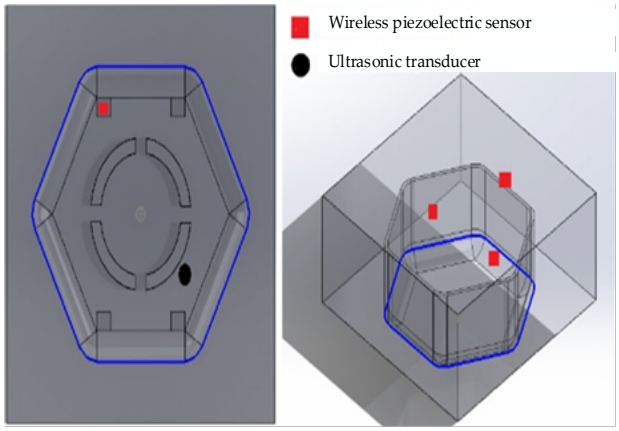

Figure 13 : Placement of the sensors inside the cavity.

4.5. Piezoelectric Sensor Measurement

Using data from the CAE simulation, we evaluate the signal transmitted by the piezoelectric sensor, Figure 14 illustrates the MATLAB block diagram. The input is the pressure of the area where sink marks are monitored. This pressure is transformed into a voltage [12]. Then noise is added to emulate real-world measurement disturbances from industrial environment.

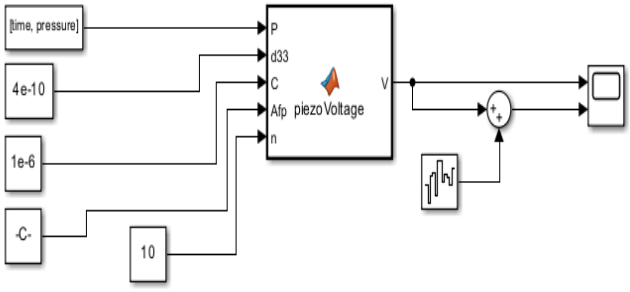

Figure 14 : Block diagram of piezoelectric sensor measurement.

Figure 15 illustrates the actual voltage while Figure 16 illustrates the measured one. We notice that the measured voltage depicts the variations of the actual voltage. By calculating the root mean square error (RMSE) and the peak error we found values of 0.0003 V and 0.0012 V respectively. This indicates that the proposed sensor accurately depicts the variations of pressure in the specified region, allowing early detection of the appearance of sink marks.

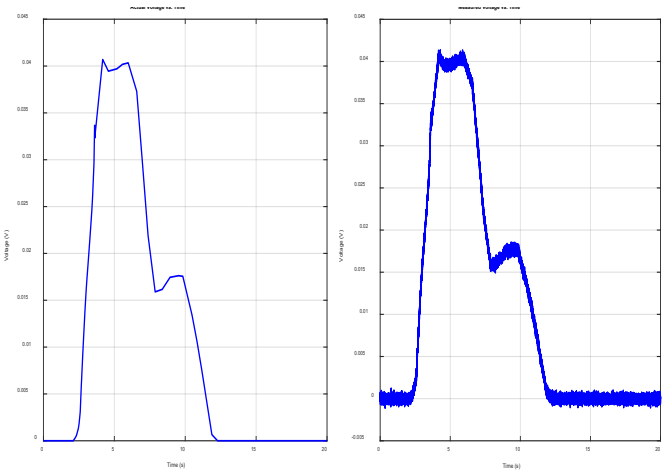

Figure 15 : Graph of the actual voltage and the measured voltage.

5. Cost Study

To assess the system's profitability and its potential impact on plastic part manufacturing, we propose a cost analysis based on a large production volume injection molding process [26]. The production cost of the part and the system's implementation expenses are shown in Table 3. Production volume, production technique, mold type and cost, lead time to final component, material cost, labor cost, and outsourcing cost are all factors that go into estimating the product's cost. Fixed costs on the other hand cover the data acquisition system's initial deployment fees.

Table 3: Detailed costs.

Variable Costs	Amount (USD)
Part cost	1.70
Production volume per month	100,000.00
Total	170,000.00
Fixed cost	
Sensor price	750.00
DAC module	108.00
Overhead costs	4,000.00
Selling expenses	1,000.00
Investment costs	5,000.00
Management expenses	2,000.00
total	12,858.00
Sales	
Part price	4.00
Total	400,000.00

To measure the effect of the system on production we implemented a breakeven analysis as seen in Figure 16. The break-even point is calculated using the following equation (5):

$$Breakeven\ point = \frac{fixed\ cost}{selling\ price\ per\ unit - variable\ cost\ per\ unit}, \quad (6)$$

The breakeven point for the suggested system was 5,590 units. Considering a monthly production and sales volume of 100,000 units, the system's implementation expenses will be recovered within the first month of adoption.

It is also important to mention that the acceptable percentage of scrab can vary between 1% to 5%. If we consider a monthly plastic production of 100,000 parts weighing 20 g, the total amount of plastic waste generated monthly will be between 20 kg to 100 kg. These values are very high considering that the cost of raw material ranges from $1 to $5 per kilogram, and the high carbon footprint of ABS estimated at 146 g CO2e/kg [27]. Therefore, successful implementation of the system will lead to a reduction of scrab, and pollution generated by injection molding.

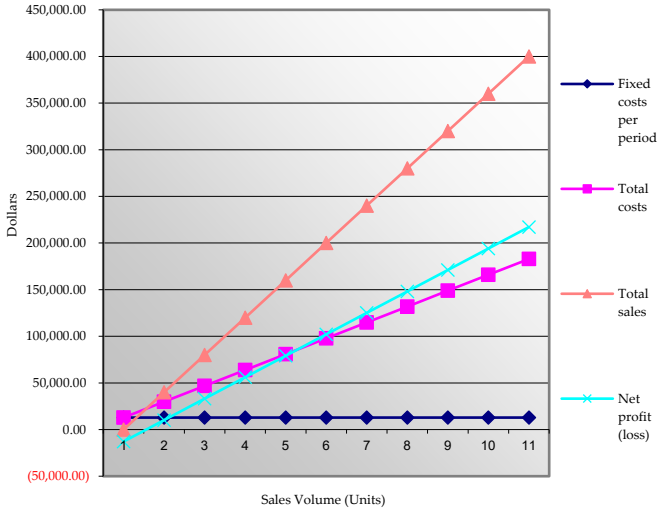

Figure 16 : Breakeven analysis chart.

6. Conclusion

In this paper, a cavity data acquisition system for in-mold process monitoring was proposed. The system allowed non-destructive measurement of cavity process parameters thanks to the use of piezoelectric and ultrasonic sensors.

The system has been developed in various phases, starting with part analysis using CAD software, followed by an investigation of defects and their causes, the sensors selected have then been described, and finally, the transmission scheme has been detailed.

Simulation results illustrated that CAD software can simulate melt behavior while allowing early detection of abnormalities and planning of sensor positioning. Similarly, piezoelectric sensors demonstrated accurate measurement of pressure showcasing that ultrasonic-based transmission of data is the best method for nondestructive monitoring of injection parameters within the cavity.

A draft cost analysis has been proposed to illustrate the cost effectiveness of the system and the short recovery of investment costs.

These results highlight the ability of the system to control injection conditions inside the mold while also improving production quality and reducing injection molding costs.

Although virtual testing provided an efficient evaluation of the system, on-site testing would allow practical investigation and examination of how the dynamic nature of the process may affect the system's capabilities. Additionally, the performance of the proposed sensors can be compared to emerging sensors such as infrared and to other commercially available sensors. Furthermore, the proposed system can be evaluated on other plastic parts to test its reliability on injection systems with varying degrees of complexity.

Future work would address the highlighted limitations and investigate the integration of a real-time control system to allow automated adjustment of in-mold parameters.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] B. P. Federation, "Plastics applications," British Plastics Federation. [Online]. Available: https://www.bpf.co.uk/plastipedia/applications/Default.aspx

[2] M. Naik, "5 ways plastics revolutionized the healthcare industry," MPO Magazine, Oct. 09, 2017. [Online]. Available: https://www.mpo-mag.com/exclusives/5-ways-plastics-revolutionized-the-healthcare-industry/

[3] K. Pulidindi and K. Ahuja, "Injection molded plastics market size, growth report – 2032," Global Market Insights Inc. [Online]. Available: https://www.gminsights.com/industry-analysis/injection-molded-plastic-market

[4] G. Berger-Weber and S. S. Aminabadi, "The injection mold as a cyber physical system: Using simulation to train its artificial intelligence," in Proc. 29th Leoben Kunststoffkolloquium, Leoben, Austria, vol. 29, pp. 1–10, Sep. 2021.

[5] J. Krantz et al., "In-mold rheology and automated process control for injection molding of recycled polypropylene," Polym. Eng. Sci., vol. 64, no. 9, pp. 4112–4127, Jun. 2024, doi: 10.1002/pen.26836.

[6] M. Baum, D. Anders, and T. Reinicke, "Enhancing injection molding simulation accuracy: A comparative evaluation of rheological model performance," Appl. Sci., vol. 14, no. 18, p. 8468, Sep. 2024, doi: 10.3390/app14188468.

[7] T. Ageyeva, S. Horváth, and J. Kovács, "In-mold sensors for injection molding: On the way to Industry 4.0," Sensors, vol. 19, no. 16, pp. 1–21, Aug. 2019, doi: 10.3390/s19163551.

[8] B. Silva, J. Sousa, and G. Alenya, "Machine learning methods for quality prediction in thermoplastics injection molding," in Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET), Cape Town, South Africa, pp. 1–6, Dec. 2021, doi: 10.1109/ICECET52533.2021.9698455.

[9] G. Gordon, D. O. Kazmer, X. Tang, Z. Fan, and R. X. Gao, "Quality control using a multivariate injection molding sensor," Int. J. Adv.

Manuf. Technol., vol. 78, no. 9–12, pp. 1381–1391, Jan. 2015, doi: 10.1007/s00170-014-6706-6.

[10] V. Vara, "The benefit of pressure sensor in injection moulding," Efficient Innovations. [Online]. Available: https://www.efficientinnovations.in/the-benefit-of-pressure-sensor-in-injection-moulding/

[11] A. Schott et al., "Development of thin-film sensors for in-process measurement during injection molding," Procedia CIRP, vol. 120, pp. 619–624, Jan. 2023, doi: 10.1016/j.procir.2023.09.048.

[12] L. Zhang et al., "A self-energized sensor for wireless injection mold cavity pressure measurement: Design and evaluation," J. Dyn. Syst. Meas. Control, vol. 126, no. 2, pp. 309–318, Jun. 2004, doi: 10.1115/1.1767850.

[13] A. S. Babalola, "Design and construction of a sensor analytic system for the monitoring of the parameters of a plastic injection mould," Front. Eng. Built Environ., vol. 1, no. 1, pp. 32–40, May 2021, doi: 10.1108/FEBE-02-2021-0003.

[14] A. Bur and C. Thomas, "An optical sensor for polymer injection molding," Electrochem. Soc. Trans., 2000. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=851715

[15] T. E. P. Gomes et al., "Development of an Open-Source Injection Mold Monitoring System," Sensors, vol. 23, no. 7, p. 3569, Mar. 2023, doi: 10.3390/s23073569.

[16] R. M. C. Bernardo et al., "Instrumentation and In-mould Data Acquisition System for Injection Moulding Process," 2023 4th Int. Conf. Artif. Intell., Robot. Control (AIRC), Cairo, Egypt, 2023, pp. 53–57, doi: 10.1109/AIRC57904.2023.10303047.

[17] O. Haberchad and Y. Salih-Alj, "A cavity data acquisition system for defect prevention in injection molding," in Proc. 10th Int. Conf. Appl. Syst. Innov. (ICASI), Kyoto, Japan, pp. 244–246, 2024, doi: 10.1109/ICASI60819.2024.10547795.

[18] B. Olmsted and M. Davis, *Practical Injection Molding*, 1st ed. Boca Raton, FL, USA: CRC Press, 2001, pp. 1–232, doi: 10.1201/9781482294590.

[19] N. Zhao et al., "Recent progress in minimizing the warpage and shrinkage deformations by the optimization of process parameters in plastic injection molding: A review," Int. J. Adv. Manuf. Technol., vol. 120, no. 1–2, pp. 85–101, Feb. 2022, doi: 10.1007/s00170-022-08859-0.

[20] S. Jiang et al., "Reducing the sink marks of a crystalline polymer using external gas-assisted injection molding," Adv. Polym. Technol., vol. 2020, pp. 1–8, Feb. 2020, doi: 10.1155/2020/3793505.

[21] M. Kariminejad et al., "Ultrasound sensors for process monitoring in injection moulding," Sensors, vol. 21, no. 15, pp. 1–22, Jul. 2021, doi: 10.3390/s21155193.

[22] V. Speranza, U. Vietri, and R. Pantani, "Monitoring of injection moulding of thermoplastics: Adopting pressure transducers to estimate the solidification history and the shrinkage of moulded parts," Strojniški Vestn. – J. Mech. Eng., vol. 59, no. 11, pp. 677–682, Jul. 2013, doi: 10.5545/sv-jme.2013.1000.

[23] K. Straka et al., "To the measurement and influences of process parameters variations on the axial melt temperature profile in the screw chamber of an injection molding machine," in Proc. SPE ANTEC, Anaheim, CA, USA, pp. 1645–1651, May 2017.

[24] B. Praher et al., "Ultrasound based monitoring of the injection moulding process: Methods, applications and limitations," AIP Conf. Proc., pp. 159–162, Jan. 2014, doi: 10.1063/1.4873755.

[25] Y. Li et al., "An optimal design method for improving the efficiency of ultrasonic wireless power transmission during communication," Sensors, vol. 22, no. 3, p. 727, Jan. 2022, doi: 10.3390/s22030727.

[26] "How to estimate injection molding cost?," Formlabs. [Online]. Available: https://formlabs.com/global/blog/injection-molding-cost/

[27] J. Tinz, T. De Ancos, and H. Rohn, "Carbon Footprint of Mechanical Recycling of Post-Industrial Plastic Waste: Study of ABS, PA66GF30, PC and POM Regrinds," Waste, vol. 1, no. 1, pp. 127–139, Dec. 2022, doi: 10.3390/waste1010010.

OUMAYMA HABERCHAD received a bachelor's degree in general engineering with a concentration in mechatronics at Al Akhawayn University in Ifrane (AUI). She is pursuing her Ph.D. at King Fahd University of Petroleum and Minerals in Systems and Control Engineering. She served as a research assistant in the Interdisciplinary research center for smart mobility and as a data analyst at the office of institutional research and effectiveness. She participated in fieldwork experiments to collect data on risk-taking in rural areas. She served in the automotive industry as a product engineer. She has published five papers. Her research interests include sustainable manufacturing, robotization, digitalization, data analysis, and geographic information systems.

YASSINE SALIH ALJ received the Bachelor's degree in microelectronics from the University of Quebec at Montreal (UQAM), Montreal, Quebec, Canada, in 2001, and the Master's degree in electrical engineering from the École de Technologie Supérieure (ETS), Montreal, Quebec, Canada, in 2003, and the Ph.D. degree in Telecommunications from the National Institute of Scientific Research – Energy, Materials & Telecommunications (INRS-Telecom), Montreal, Quebec, Canada, in 2008. He served as a research assistant at the Telebec Underground Communications Research Laboratory (LRTCS) from 2005 to 2008, and then during 2009 as a Postdoctoral Fellow at Poly-Grames Research Center, of the École Polytechnique de Montréal, Montreal, Quebec, Canada. He is currently working as a Full Professor of Engineering at the School of Science and Engineering (SSE) of Al Akhawayn University in Ifrane (AUI), Morocco, where he also served during 2019-2022 as Academic Coordinator for General Engineering programs (GE and RESE majors - Renewable Energy Systems Engineering). He has published over 80 publications and has been actively involved in IEEE events for the past five years, where he chaired and served as Technical Program Member or as distinguished

reviewer for over 200 conferences. His research interests are in the areas of Wireless Communications, Indoor Positioning, UWB (Ultra-Wideband), Smart Systems, GPS (Global Positioning System) and Engineering Education.

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

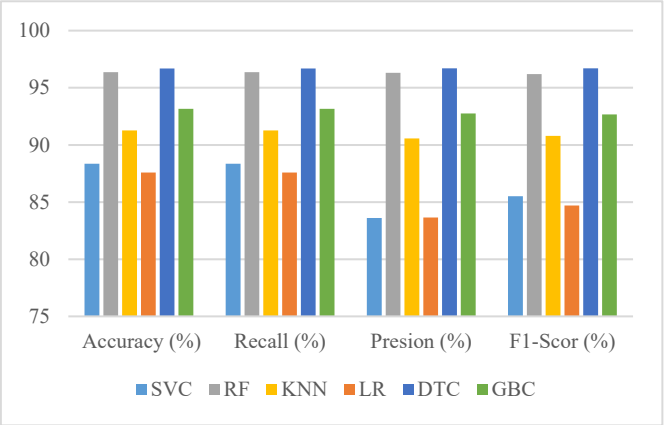

Figure 2: Performance Plot of Proposed Models

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.