ABUAD Journal of Engineering Research and Development (AJERD)
ISSN (online): 2645-2685; ISSN (print): 2756-6811

Volume 8, Issue 3, 94-116

Lightweight CNN Architectures for Fault Diagnosis of Power Generator sets: A Comparative Study of MobileNet and AlexNet

Ekerette Bernard IBANGA, Kingsley Monday UDOFIA, Kufre Michael UDOFIA, Unwana Ubong IWOK, Emmanuel Oluropo OGUNGBEMI

Department of Electrical and Electronics Engineering, University of Uyo, Uyo, Nigeria
ebibanga@gmail.com/kingsleyudofia@uniuyo.edu.ng/kufreudofia@uniuyo.edu.ng
/unwanaiwok@uniuyo.edu.ng/ogungbemioluropo@uniuyo.edu.ng

Corresponding Author: unwanaiwok@uniuyo.edu.ng, +2348069592778
Received: 17/07/2025
Revised: 09/09/2025
Accepted: 17/10/2025
Available online: 25/10/2025

Abstract: *The use of power generator sets (3.5kVA – 5.5kVA) for domestic and commercial backup supply has become a mainstay in Nigeria due to the unstable electricity from the grid. To keep these backup supplies running, traditional diagnostic approaches that are reliant on manual inspections and physical measurements have been adopted. These are often time-consuming, reactive, and unsuitable for real-time monitoring. To address these challenges, a machine learning approach is presented by performing a comparative analysis of MobileNet and AlexNet convolutional neural networks for automated audio-based fault diagnosis in 5kVA generators. Fault signatures are obtained from acoustic data recorded from 25 generator units under five operational states—Normal, Caburetter, Exhaust, Valve, and Plug faults. Mel-Frequency Cepstral Coefficients (MFCC), Continuous Wavelet Transform (CWT), and Short-Time Fourier Transform (STFT) were employed to transform the raw audio signals into two-dimensional spectrograms that contain both temporal and spectral fault signatures. Using transfer learning, these spectrograms were utilized as input features to train versions of MobileNet and AlexNet, which were pre-trained on ImageNet weights. Their performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Results obtained from the evaluation metrics show that MobileNet significantly outperformed AlexNet across all feature transformations (MFCC, CWT, and SSTFT). It achieved a peak accuracy of 92% and an AUC of 0.99 with STFT spectrograms. In contrast, AlexNet achieved lower accuracies (54–59%), indicating lower discriminative power. The class-wise ROC-AUC analyses confirmed that MobileNet achieved near-perfect classification, particularly in distinguishing between Normal and any of the fault conditions, while AlexNet struggled with subtle classes, such as Plug and Valve faults. These findings indicate that STFT is the most discriminative spectrogram and MobileNet is the best-performing diagnostic framework. This makes it suitable for deployment in resource-constrained environments and edge devices. This research contributes to the advancement of intelligent, real-time condition monitoring of domestic generator sets, thereby reducing downtime and enhancing energy reliability in off-grid contexts.*

Keywords: *MobileNet, AlexNet, Fault Diagnosis, Feature Transformation, Power Generators, Audio Database*

1. INTRODUCTION

The epileptic and persistent inadequacy of the national power supply has led to widespread dependence on small-scale power-generating sets (SSPGS). The advantage of balancing affordability, portability, and capacity has made the 3.5kVA to 5.5kVA range power generator sets (PGS) the most widely used ones for domestic and small businesses [1]. These generators serve as critical backup power sources, bridging the gap between demand and the unreliable public electricity grid. However, these generators are plagued by operational faults, including carburetor, exhaust, valve, and plug failures that occur due to poor maintenance practices, fuel quality issues, high speed, heavy load, and extended use under poor conditions [2], and repeated run cycles over time subject the generator to wear with a consequent loss of technical efficiency. If left undiagnosed, such faults can lead to performance degradation, increased fuel consumption, or a complete breakdown, resulting in a significant loss of turnover for the business. For this reason, it is crucial to predict failures in advance, allowing for the replacement of faulty parts before they reduce machine performance.

Preventive identification of parts failures in PGS plays a crucial role in machine maintenance. While identifying faults is a key step in protecting against system collapse, early detection prevents failures from becoming serious problems [3]. The traditional fault diagnosis method is based on a variety of different signal acquisition and signal processing means for equipment fault diagnosis and detection. For PGS in particular, fault diagnosis is established on engineering principles, signal analysis, and physical inspections, typically involving manual inspections or the intervention of technicians after a failure has occurred. Examples of some of these traditional methods include Vibration Analysis [4], Thermal/Infrared Imaging [5], Acoustic Emission Analysis [6], Current and Voltage Signature Analysis [2], Oil Analysis (for lubricated generators) [7], Visual Inspection [8], Partial Discharge Testing [9], and Electromagnetic Signature Analysis [10]. These reactive approaches, while effective in large-scale industrial machines, are impractical for small generator users due to

their reliance on expert knowledge and diagnostic rules, emphasizing the understanding of underlying physical phenomena and using direct observations and measurements to identify potential issues. They are not only time-consuming and costly but also fail to prevent the long-term degradation of the PGS. The need for automated, real-time fault detection systems that can operate effectively in domestic settings with limited technical expertise and resources has led to a paradigm shift toward data-driven and machine learning-based methods [11].

Research advancements in machine learning have made intelligent systems capable of early fault detection and classification increasingly viable. The conventional machine learning (ML) techniques, such as Support Vector Machine (SVM) [12], Random Forest (RF) [13], k-Nearest Neighbours (k-NN), and Naïve Bayes (NB) [14] amongst others, have been increasingly used for fault diagnosis in power generator systems due to their ability to learn patterns from handcrafted features extracted from data and detect faults more timely compared to traditional methods. These handcrafted features extracted from sensor data, though useful, may struggle with growing complexity and rapid evolution of modern software requirements, posing limitations such as limited adaptability, data inconsistencies, limited coverage, and the needed accuracy to handle complex, nonlinear patterns in system behaviour [3].

Currently, deep learning-based methods are considered the most effective due to their ability to automatically and efficiently extract salient features from images and provide end-to-end fault diagnosis without the need for explicit feature extraction or complex image processing and manual feature engineering [15]. Recent advances in deep learning have developed systems for the fault detection and diagnosis using a combination of low-cost sensors and deep learning algorithms, such as Artificial Neural Network (ANN) algorithms, Convolutional Neural Network (CNN) model, Recurrent Neural Network (RNN) applications [16], and Deep Generative Systems. This approach offers promising alternatives for analysing high-dimensional data by directly learning fault patterns from raw sensor data, with the capacity to generalize models across different scenarios, identifying subtle anomalies, and classifying various fault types in mechanical and electrical systems. However, they are computationally intensive, memory-demanding, and unsuitable for deployment on mobile or edge devices, which limits their practicality in resource-constrained environments [17]. There is also the paucity of comparative studies evaluating multiple deep learning models under the same conditions for SSPGS fault diagnosis, and the reliance on simulated and standard fault datasets raises questions about generalizability to real environments. These gaps have motivated the exploration of alternative deep learning frameworks with reduced complexity and memory requirements for resource-constrained environments, as well as the generation of customized datasets for model development.

To address these limitations, this work proposes using lightweight convolutional neural network architectures, AlexNet and MobileNet, in resource-constrained environments as a promising solution for fault diagnosis of 5kVA SSPGS. The performance of the MobileNet and AlexNet deep learning architectures in classifying the faults of 5kVA SSPGs using customised acoustic data is evaluated comparatively using commonly used performance evaluation metrics. The customised audio signals from the 5 kVA SSPGS are transformed into spectrograms using three feature transformation techniques- Mel-Frequency Cepstral Coefficients (MFCC), Continuous Wavelet Transform (CWT), and Short-Time Fourier Transform (STFT) - and then used as inputs to both models. These spectrograms contain the acoustic signatures under normal and common fault conditions for automated audio-based fault diagnosis in 5kVA SSPGS.
The expected outcome is to identify the most effective model–feature combination that achieves high diagnostic accuracy. This demonstration will lay the groundwork for an automated intervention that is scalable, cost-effective, and timely. It will have a significant socioeconomic impact and enhance the reliability and sustainability of small-scale off-grid power solutions.

Numerous research studies have been conducted on fault detection and diagnosis for rotary machines, wind turbines, induction motors, automobile and aircraft engines, and other heavy-duty industrial machines. These studies use simulated data and multiple monitoring sensors, demonstrating their applicability in practical engineering scenarios. However, our research has not been able to uncover any substantial research on common fault diagnosis for domestic and small-scale commercial power generating sets (DaSSC-PGS) within the 3.5 kVA to 5.5 kVA capacity range. Furthermore, despite the growing application of machine learning, signal processing techniques, and the strong potential of deep learning methods for fault diagnosis, the most commonly used techniques to maintain the backup supply of these domestic and small-scale commercial DaSSC-PGS are traditional diagnostic approaches that rely on manual inspections and physical measurements. There has been no research, to the best of our knowledge, on the use of machine learning approaches for fault diagnosis or any comprehensive comparative evaluation of different models under varying fault conditions for this category of generators.

Therefore, this study systematically investigates the performance of two deep learning architectures - MobileNet and AlexNet - using customised datasets of four common fault conditions (Caburettor, Exhaust, Valve, and Plug faults) and one Normal operating condition. The comparative analysis under each of the operating scenarios provides an understanding of the models' effectiveness in realistic engineering scenarios, offering useful insights into the models' behaviour.

The remainder of this manuscript is organized as follows. The following section presents the literature review. Section 3 describes the methodology of the proposed fault diagnostic system, based on DaSSC-PGS data acquisition and transformation, as well as model development. Section 4 presents the obtained results, along with a discussion. The final section presents the paper's conclusions and recommendations.

2. LITERATURE REVIEW

For instance, Shang et al. [1] developed a transformer fault diagnosis model based on dissolved gas analysis, where the multi-scale approximate entropy of gas concentrations was calculated under different fault scenarios. These entropy features were used to train a convolutional neural network (CNN), which was subsequently optimized using an enhanced sparrow search algorithm, yielding improved diagnostic performance. The major contributions of the work were the proposed ISSA-optimized CNN, which reduced memory consumption, achieving high diagnostic accuracy and efficiency. With F1-scores consistently above 85% across fault types, the method shows strong generalization ability, providing a robust diagnostic framework that can be extended for real-world transformer monitoring. The limitations of the work, however, stem from the data, as the study utilized a limited DGA dataset. The authors acknowledged the need to collect more on-site transformer fault data to validate the model's practicality.

Building on the integration of deep learning in fault detection, Khaleefah et al. [18] proposed a Long Short-Term Memory (LSTM) model for early fault identification in electrical power transmission networks. The model processed three-phase current and voltage signals from a single terminal to detect faults with high precision. With a detection accuracy of 99.65% and an error rate of only 1.17%, this approach outperformed conventional neural networks (93.55%) and CNN-based models (94.60%), demonstrating its robustness in enhancing grid reliability. The key contribution of the paper is the development of a high-performing LSTM-based model for fault detection in transmission systems, which achieves near-perfect accuracy compared to traditional ANN and CNN methods. However, gaps remain due to the limited scope of fault types, mainly focused on line-to-ground faults. Other faults (line-to-line, double-line-to-ground, three-phase symmetrical faults) were not fully included, and the dataset limitations rely on a MATLAB/Simulink simulated dataset. Real-world fault data is needed to validate generalization and practical applicability.

To improve power quality disturbance (PQD) detection, Yoon and Yoon [19] introduced a novel voltage signal segmentation method as input to transformer-based deep learning models. Synthetic voltage signals were generated under four fault conditions using the IEEE 9-bus system simulated in PSCAD/EMTDC. These segmented signals were then classified by deep models, which successfully identified both the type and location of disturbances in real-time, further underscoring the efficacy of advanced preprocessing and modelling techniques in power systems. The paper makes significant contributions by proposing and validating a real-time fault detection framework. This framework integrates signal processing tools and intelligent decision-making models into a hybrid system, offering improved accuracy and speed compared to traditional techniques. One of the gaps in the work is that the evaluation was based on simulation results; real-world deployment and field validation are not fully addressed.

Extending the application of deep learning to mechanical systems, Nashed et al. [4] presented a real-time fault classification approach for a lab-scale gas turbine using acoustic emission (AE) signals. Time-frequency features obtained via continuous wavelet transform, along with statistical features such as RMS and kurtosis, were fused into RGB images to train a CNN. The model effectively distinguished between normal and faulty conditions across varying turbine speeds, enabling automated, real-time condition monitoring with minimal human oversight. The main contribution of this work is the development of a novel online monitoring system, which utilizes acoustic emissions (AE) to detect and classify normal and faulty operating conditions in gas turbines, both with and without load. One of the gaps identified in the work is that the developed model is not suitable for resource-constrained environments, such as mobile devices and edge-embedded systems.

Similarly, Kowalski et al. [5] proposed an intelligent fault diagnosis system for a stroke diesel marine engine using an automated machine learning approach in decomposition mode. Engine-emitted signals were processed through a one-vs-one classification scheme involving Extreme Learning Machines (ELMs), with Error-Correcting Output Codes (ECOC) used to reconstruct the multi-class problem. The system achieved high accuracy and fast inference time on real-world data, highlighting its potential for onboard, real-time fault detection in maritime applications.

Further expanding the scope of signal-based diagnostics, Mitiche et al. [6] explored the use of electromagnetic interference (EMI) signals for identifying insulation and mechanical faults in high-voltage (HV) electrical systems. A two-stage 1D-CNN architecture was proposed, utilizing transfer learning to first filter relevant signals and then classify fault types directly from raw time-domain data. This method reduced computational overhead and achieved high accuracy, with successful deployment in industrial HV monitoring instruments validated under both lab and field conditions.

From the numerous studies conducted in the area of fault detection and diagnosis, our review has revealed that the focus has been on heavy-duty industrial rotary machines, wind and gas turbines, induction motors, automobiles, marine and aircraft engines, utilizing simulated or online generic data. Additionally, a significant gap exists in the commonly used DaSSC-PGS within the 3.5kVA to 5.5kVA capacity range, as these have not been explored for machine learning-based fault diagnosis solutions. Furthermore, we have not found any comparative evaluation of machine learning models to assess their behaviour under common fault conditions associated with this category of generators. This has motivated this work to address the 5kVA category of DaSSC-PGS using deep learning architectures trained on customised data obtained from the on-site PGS.

3. METHODOLOGY

This section outlines the methodological framework of the developed fault diagnosis and characterization system, utilizing deep learning. A combination of hardware and software resources was utilized to achieve this work, which is presented as follows:

3.1 Materials

The list of materials, both software and hardware, used and their specifications are as follows:

i. Infinix Smart 8 mobile phone with HIOS 7.6; Chipset: MediaTek 6762 Hello P22 (12nm); CPU: Octa-core 2.0 GHz Cortex-A53; GPU: PowerVR GE8320.
ii. HP Laptop running Windows 10 Pro, 11th Generation Intel® Core™ i5 processor, 16 GB memory; Intel® Iris® Xᵉ Graphics.
iii. Jupyter Notebook Python software with a suite of specialized libraries, including Scikit-learn, Librosa, Scipy, Spafe, Noise Reduce, and Pywavelets, was used for audio signal processing, data pre-processing tools including Pandas and NumPy, and Visualization tools like Matplotlib and Seaborn.

3.2 Method

The approach for the proposed deep learning-based fault diagnosis and classification system for the 5kVA DaSSC-PGS consists of the following stages: data collection, data preprocessing, feature transformation and extraction, model development, and model evaluation. In the data collection and description, data preprocessing, feature transformation, and extraction stages, signal processing techniques such as Mel Frequency Cepstral Coefficients (MFCC), continuous wavelet transformation (CWT), and Short-Time Fourier Transform (STFT) are used to convert the data from signal formation into spectrograms. The second stage of the proposed algorithm focuses on developing ImageNet-pretrained MobileNet and AlexNet through transfer learning, using data samples collected for the different classes of faults: Carburettor, Exhaust, Valve, Plug, and Normal, as discussed in sections 3.2.1 and 3.3. The idea of transfer learning in Deep Learning is to reuse knowledge that the previously learned model derived from a particular learning problem. This approach enhances the learning process by setting the starting point of a model on a related problem [6]. The two models were evaluated to identify the best-performing model. All the stages mentioned in the above workflow are discussed in the subsequent subsections. The structured methodology used to develop the fault diagnosis and characterization system is presented in Figure 1.

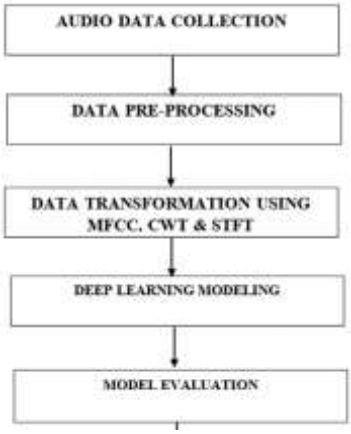


Figure 1: Block diagram of system methodology

3..2.1 Data collection and description

The sound of a machine contains information about its condition, and the presence of a fault can be detected from the machine's acoustic characteristics. The ability to extract this fault signature from the generator's sound is a handy tool in fault diagnosis. The audio signals, ranging from 0 to 20 kHz, were collected using a phone microphone. A non-intrusive approach was employed by positioning the microphones 2-10 cm from the desired source. The acoustic data was collected from 25 5kVA running generators for each of the fault conditions described below.

i. Carburetor fault caused by dirt, fuel-related issues, air-fuel mixture problems, wear and tear, and improper maintenance.
ii. Exhaust fault caused by blockages, leaks, corrosion, or damaged components like the muffler or exhaust pipe, leading to poor performance, increased emissions, or excessive noise.
iii. Valve faults in a generator can result from improper timing, wear and tear, carbon buildup, or poor lubrication, leading to loss of compression, reduced power, and engine misfires.
iv. Plug faults in a generator, such as fouling, wear, or improper gap, can lead to weak or no spark, causing difficulty in starting, misfires, or poor engine performance.

The phone's microphone was used to record the different acoustics for each fault condition, with an average duration of 90 seconds. The collected recordings were transferred to the HP laptop and carefully saved in respective folders according to the four fault conditions: Exhaust, Plug, Carburettor, and Valve. Additionally, a "Normal" category was created to indicate the normal operating condition of the generators, and data was collected accordingly. The size of the data collected

was Carburettor (3108 audio files), Exhaust (3108 audio files), Valve (5699 audio files), Plug (2849 audio files), and Normal (259 audio files). All audio signals were broken up into 100-millisecond segments.

3.2.2 Data pre-processing

The pre-processing stage involved cleaning the data by trimming out silence and removing noisy sections of the audio signals using Audacity software.

3.2.3 Data transformation

MobileNet is a CNN framework that processes image data. The step of data transformation was performed to convert the pre-processed audio signals from 1D to a rich 2D image format suitable for CNNs. This 2D image can be a spectrogram or a scalogram, depending on the transformation tool used. In this work, MFCC, STFT, and CWT were used for data transformation.

i. **Continuous wavelet transform (CWT):** CWT is a mathematical tool used to analyse signals in both time and frequency domains simultaneously. Unlike the Fourier Transform, which only provides frequency information, CWT offers a time-frequency representation that reveals how signal components vary over time. It achieves this by convolving the signal with scaled and shifted versions of a mother wavelet. The continuous wavelet decomposition of a signal x(t) is mathematically represented by the integral operation as shown in Equation 1 [8].

$$x_w(a,b) = \frac{1}{|a|^{\frac{1}{2}}}\int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \quad (1)$$

CWT is a powerful signal processing technique with better adaptability to non-stationary (transient) signals. It effectively detects localized features such as spikes, edges, or singularities. The key feature of CWT is its multi-resolution analysis capability. This provides good time resolution for high frequencies and good frequency resolution for low frequencies. CWT can adapt the window automatically based on scale. It is also continuous in both time and scale.

The mother wavelet, denoted as $\psi(t)$, is a continuous, integrable function defined across both temporal and spectral domains. Its complex conjugate, represented by ψ^*, facilitates the generation of derived wavelet functions—referred to as daughter wavelets—through scaling ($a \in \mathbb{R}^+$) and translation ($b \in \mathbb{R}$) operations. These daughter wavelets serve as localized basis functions for multi-resolution analysis, enabling the decomposition of signals into time-frequency representations that strike a balance between temporal and spectral resolution.

The Morlet wavelet shown in Equation 2 [8] combines a complex exponential carrier and a Gaussian envelope, defined by centre frequency f_0 and bandwidth σ. It offers adjustable time-frequency resolution across scales, with scale inversely related to frequency.

$$w(t) = \frac{1}{K\sigma}e^{-(\sigma t)^2}\cos(2\pi f_0 t) \quad (2)$$

The output is a scalogram of the transformed data, which serves as the input to the MobileNet model.

ii. **Mel-frequency cepstral coefficients (MFCC):** MFCC mimics the human auditory perception frequency (Mel scale) to transform an audio signal into its representative short-term power spectrum. It captures time-frequency information, similar to a spectrogram. The pipeline for the MFCC data transformation is shown in Figure 2.

iii. **Pre-emphasis**: The signal was emphasized to amplify high frequencies, which often contain important information. Pre-emphasis was implemented using the mathematical expression in Equation 3 [9].

$$s(t) = x(t) - \alpha x(t-1) \quad (3)$$

where α is a dimensionless factor typically about 0.97, and x(t) is the original signal.

iv. **Framing and windowing**: The data was framed into 100-millisecond segments, and a windowing overlap of 30% was done to account for discontinuities that occur between adjacent segments due to framing and to prevent loss of vital information. Each overlapping frame contains a short signal segment that captures local frequency content. The frames are defined according to Equation 4 [9]

$$x_m(t) = x(t + mH), 0 \le t < N \quad (4)$$

Where $x_m(t)$ represents the mth window, m is the frame index, and H is the hop size.

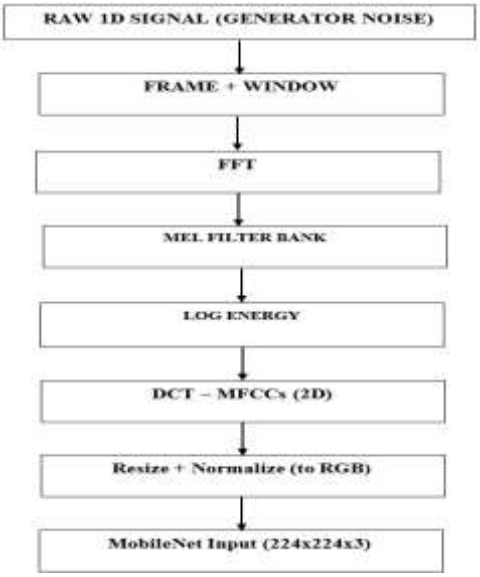


Figure 2: MFCC pipeline for data transformation to spectrogram

The Hamming window function was applied to each of the segmented frames to mitigate spectral leakages. The Hamming window function is expressed by Equation 5 [9].

$$w(t) = 0.54 - 0.46\cos\left(\frac{2\pi t}{N-1}\right) \quad (5)$$

The window size is equal to the frame size N

v. **Fast fourier transform (FFT):** The Fast Fourier Transform (FFT), as shown in Equation 6 [20], was applied to each windowed temporal segment of the audio signal to derive its spectral representation. This approach enables time-localized frequency decomposition, restructures the DFT matrix into sparse, simplified components, and enhances computational efficiency by drastically reducing the algorithmic complexity and enabling rapid frequency analysis.

$$X_m(k) = \sum_{n=0}^{N-1} x_m(n) \cdot e^{-i\frac{2\pi kn}{N}} \quad (6)$$

Where $X_m(k)$ is the FFT of the mth frame at the frequency bin k, i is an imaginary unit, and N is the number of points in the FFT.

vi. **Filter-bank operation:** Mel filter banks (a perceptual scale representing higher frequencies with lower resolution) consist of overlapping bandpass filters with triangular magnitude responses, strategically positioned across the signal's spectral energy spaced according to the Mel scale. The overlap ensures smooth transitions between frequency bands. Mel filter banks utilize the psychoacoustic properties of the Mel scale to approximate the nonlinear frequency resolution of the human auditory system, exhibiting finer resolution at lower frequencies and logarithmic spacing at higher frequencies. The Mel scale is defined as shown in Equation 7 [9].

$$m(f) = 2595.\log_{10}\left(1 + \frac{f}{700}\right) \quad (7)$$

Where f is the given frequency in Hz.
The filters are defined such that:
The filter starts at a lower frequency f_i
Peaks at a center frequency f_c
Ends at an upper frequency f_{i+1}

vii. **Logarithm of energy:** multiply the power spectrum by each filter bank and sum the result to obtain the total energy in each Mel band as shown in Equation 8 [10].

$$E_m = \sum_{k=f_i}^{f_{i+1}} P(k) \cdot H_m(k) \quad (8)$$

Where E_m is the energy in the mth Mel filter, $\mathcal{P}(k)$ is the power spectrum at frequency k, and $H_m(k)$ is the Mel filter's magnitude response of the mth filter at frequency k.

The log-transformed power spectrum is converted into cepstral coefficients through the inverse Fourier transform, effectively decorrelating the features to enhance their suitability for pattern recognition and machine learning models. This process leverages the mathematical properties of the cepstral domain to isolate independent signal components, thereby optimizing discriminative performance in applications such as speech or audio analysis. This is achieved using the Discrete Cosine Transform (DCT), which serves as a type of inverse Fourier transform. The DCT, shown in Equation 9 [9], is used to convert the log-power spectrum from the spectral to the cepstral domain. The DCT helps in compressing most of the signal information into the first few coefficients, making the features more compact and easier to model.

$$C_n = \sum_{m=0}^{M-1} \log \text{Energy}_m \cdot \cos\left[\frac{\pi n}{M}\left(m + \frac{1}{2}\right)\right] \quad (9)$$

Where C_n denotes the n-th Mel-Frequency Cepstral Coefficient, capturing spectral characteristics in the cepstral domain. M represents the total number of triangular Mel-spaced filters applied to the power spectrum, and n defines the truncation limits for the cepstral coefficients. $\log \text{Energy}_m \cdot q$ Quantifies the logarithmic energy output of the mm-th Mel-frequency band, enhancing perceptual relevance.

viii. **Short-Time Fourier Transform (STFT):** The Short-Time Fourier Transform (STFT) signal processing technique was used to analyse the frequency content of the acoustic signals, extracting time and frequency information from the windowed signal. The expression for STFT is shown in Equation 10 [21]

$$\text{STFT}(t, f) = \int_{-\infty}^{\infty} s(\tau) h^*(\tau - t) e^{-j2\pi f\tau} d\tau \quad (10)$$

Where $s(\tau)$ is the original continuous-time signal; $h*(\tau - t)$ represents the complex conjugate of a time-localized windowing function $h(\tau - t)$ centered around time t. This windowing operation extracts the localized portion of the signal around t, enabling time-resolved spectral analysis; e^{-i2nft} is the complex exponential representing the Fourier Transform, where j represents the imaginary unit; t is the time parameter, and f is the frequency parameter; τ is the integration variable.

The resolution of the STFT was determined by the frame size N and the hop size (or step size) H. Frame size (N): 100ms (4410 samples), Hop size (H): 30% overlap ($0.7 \times H$) from Equation 11 [21]. The STFT can be visualized as a spectrogram and serves as input to CNNs for audio-based models.

$$x_m(t) = x(t + mH), 0 \leq t < N \quad (11)$$

Where $x_m(t)$ represents the m-th window, m is the frame index, H is the hop size

3.3 Development of Deep Learning Models

The methodology utilized pre-processed and CWT-transformed images, containing temporal and frequency information, as inputs to the CNN, where relevant features were extracted.

3.3.1 Convolutional neural networks (CNN)

Convolutional Neural Networks (CNNs), a category of neural network architectures, are a type of feed-forward artificial neural network whose connectivity structure is inspired by the organization of the animal visual cortex. It is optimized for analysing structured grid-based data (digital images or time-series signals), and it learns hierarchical representations of image data through a series of convolutional, pooling, and fully connected layers. In this study, two CNN variants—AlexNet (a pioneering architecture for image classification) and MobileNet (designed for efficiency on mobile platforms) are implemented to evaluate their efficacy in feature extraction and computational efficiency.

3.3.2 MobileNet

MobileNet comprises a collection of streamlined neural network models optimized for deployment on resource-constrained platforms like mobile devices and embedded vision systems. MobileNet uses depth-wise convolution and pointwise convolution (1x1 convolution) as shown in Figure 3. This dramatically lowers parameter counts and computational demands compared to conventional convolutional architectures. It also utilizes the width multiplier parameter to regulate the number of channels and the resolution multiplier parameter to adjust the input image resolution. This design makes a trade-off of computational cost and memory usage over accuracy, and because of its lightweight and optimization for speed, it is able to perform detection/diagnosis tasks efficiently.

In this work, the MobileNet algorithm for fault diagnosis was developed in segments. The first segment involved importing the MobileNet and necessary Keras components to build and train the CNN. ImageDataGenerator was used for pre-processing and loading image data from directories; Scikit-learn metrics for evaluating performance, and Matplotlib for plotting and visualization.

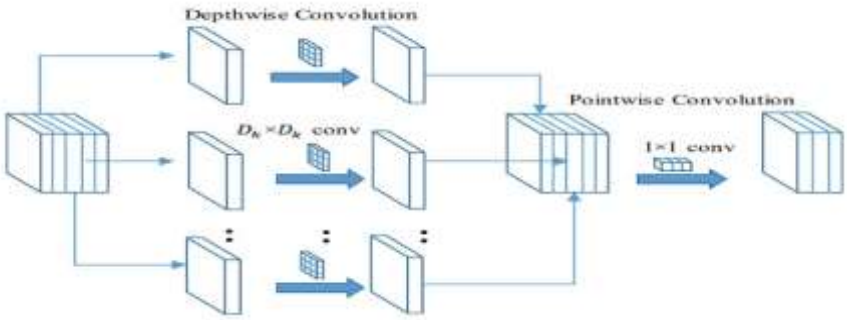


Figure 3: MobileNet architecture showing alternating depth-wise and point-wise convolution.
source: [11]

The next stage was loading the pre-trained MobileNet, which is the MobileNetV1 pre-trained on ImageNet. All MobileNet layers were set to freeze to prevent them from being updated during training. Then the customized classification layer was added for training. The snippet of code below was used to reduce the feature map dimensions using the Global Average Pooling layer.

```
num_class = 5  # Number of fault categories
x = base_model.output
x = GlobalAveragePooling2D()(x)
x = Dense(1024, activation='relu')(x)
predictions = Dense(num_class, activation='softmax')(x)
model = Model(inputs=base_model.input, outputs=predictions)
```

The fully connected layer with 1024 ReLU units was added, and then the final model, which combines MobileNet and custom layers, was created. The softmax function for multi-class classification. The helper function, def count_samples_per_class(dataset), was used for debugging class imbalance by counting the number of samples per class in a dataset. The Adam optimizer used for compiling was set with a learning rate = 0.0001. “Categorical Crossentropy” Loss was used for the multi-class classification, and the accuracy metric for tracking the performance.

```
model = Model(inputs=base_model.input, outputs=predictions)
model.compile(optimizer=Adam(learning_rate=0.0001), loss='categorical_crossentropy', metrics=['accuracy']);
```

The data was normalized [0 1], and the images resized to 224×224 to match MobileNet input requirements. The model is defined with input from MobileNet and output from the new layers. The input to MobileNet is the RGB input images, which are scalograms obtained from the data transformation stage in the image directory. The training and validation images are read from their respective folders, and class labels are automatically inferred from the subfolder names.

```
history = model.fit(
    train_dataset,
    steps_per_epoch=len(train_generator),
    epochs=20,
    validation_data=validation_dataset,
    validation_steps=len(validation_generator)
)
```

The fit is used to train the model for 20 epochs. At the end of training and evaluation, the model was saved.

3.3.3 AlexNet

AlexNet architecture, shown in Figure 4, is one of the earliest CNN architectures known for its breakthrough in large-scale image recognition. It consists of five convolutional layers, each designed to extract increasingly complex feature dynamics from the input. The first layer captures simple features like edges, while the deeper layers capture more abstract features like object parts and patterns. To reduce computational load, AlexNet employs overlapping max pooling after some of the convolutional layers. The AlexNet framework is relatively large for classification tasks, but it was chosen for its simplicity.

This work trains an AlexNet-based facial recognition system using PyTorch, with data augmentation**,** transfer learning, and evaluation. ReLU (Rectified Linear Unit) activation was used in AlexNet to learn complex patterns, speed up the training process, and extract features within the data [13]. The dropout layers were used to prevent overfitting by randomly setting a fraction of the input units to zero during training. Additionally, extensive data augmentation techniques like image translations, reflections, and patch extractions were used to ensure that the model generalizes better to new data. Proper feature extraction from the images was achieved through multiple convolutional and pooling layers, and the max-pooling feature allowed for down-sampling, spatial dimension reductions, and the extraction of the most important features.

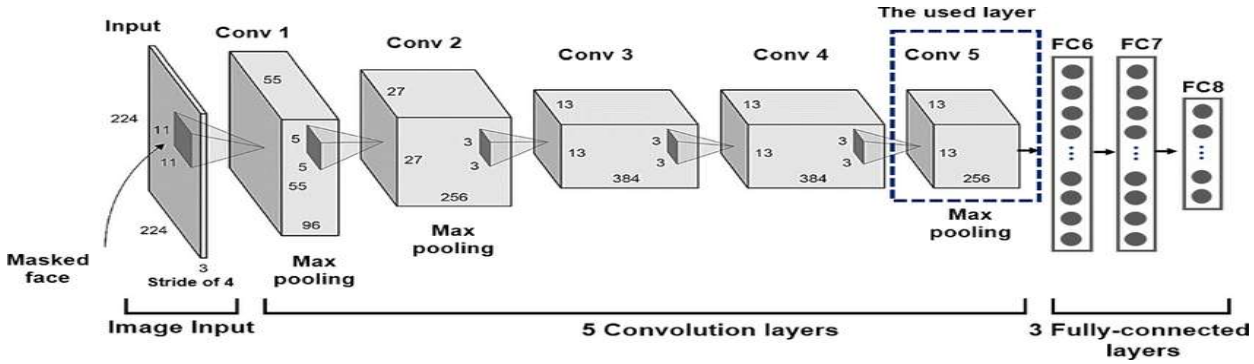


Figure 4: Facial recognition AlexNet architecture
Source: [12]

The AlexNet model was initialized by loading it with pre-trained weights (ImageNet) using the function

alexnet = models.alexnet(pretrained=True);

All the layers were frozen to retain pre-trained features except for the final fully connected layer, which was replaced to match the number of output classes (in this instance, fault classes). The LogSoftmax layer was added for classification. The Adam optimizer was used as an efficient weight update strategy.

optimizer = optim.Adam(alexnet.parameters();

At the training and validation stage, the training loop was done in multiple epochs (20 epochs), and the data was forward-propagated through the model. The calculated loss was used to update weights by gradient descent and backpropagation, while the training and validation losses and accuracy were calculated for feedback and analysis. This is useful for history tracking.

3.4 Model Evaluation

To assess the performance of the proposed classification models, the following metrics were used: accuracy, precision, recall, and F1-score. In addition, the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics were adapted for the multiclass classification with 5 classes: Exhaust, Normal, Valve, Carburetor, and Plug.
Some of these metrics are defined below:

ROC curve: The ROC Curve is a one-vs-all approach, where each line on the graph represents the model's performance in distinguishing between a specific fault class and all other classes. The curve displays the true positive rate (detection rate) plotted against the false positive rate, which is the ratio of negative instances that are incorrectly classified as positive, i.e., the specificity. The ROC results show the trade-off between the detection rate (sensitivity) and the specificity, and are particularly useful when class imbalance exists and the costs of false positives and false negatives differ. The closer the curve follows the left-top borders of the plot, the more accurate the test. Similarly, the closer the curve is to the diagonal of the plot, the less accurate the test. For the multiclass classification case in this work, the class with the highest probability is typically chosen as the decision threshold using the following formula: class = argmax(softmax output). It provides insight into how well the model distinguishes between the positive and negative classes at all classification thresholds. The AUC provides a scalar value that summarizes the entire ROC curve. It represents the likelihood that the model ranks a random positive instance higher than a random negative one. An AUC scalar value of 1.0 - Perfect model, 0.5 - no better than random guessing, and < 0.5 - Worse than random (inverted model). The AUC is threshold-independent because it evaluates performance across all possible thresholds. It is excellent for comparing models regardless of the chosen decision threshold. It also provides robustness in Model Comparison by selecting among different deep learning architectures or training configurations, and it is useful for threshold selection in classification.

AUC: measures the ability of a model to distinguish between positive and negative classes across different classification thresholds by measuring the area underneath the ROC curve. The higher the AUC, the better the model's performance in distinguishing between positive and negative classes. AUC is useful for imbalanced datasets and is used to interpret the performance of a model: a perfect model has an AUC = 1, while an AUC = 0.5 indicates the model is randomly guessing. AUC is executed using the scikit-learn library in Python using the function ROC AUC_score. Combining the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) demonstrates how well the models discriminate, regardless of data imbalance. It also prevents cases where accuracy scores can be misleading, with a high accuracy value but zero usefulness.

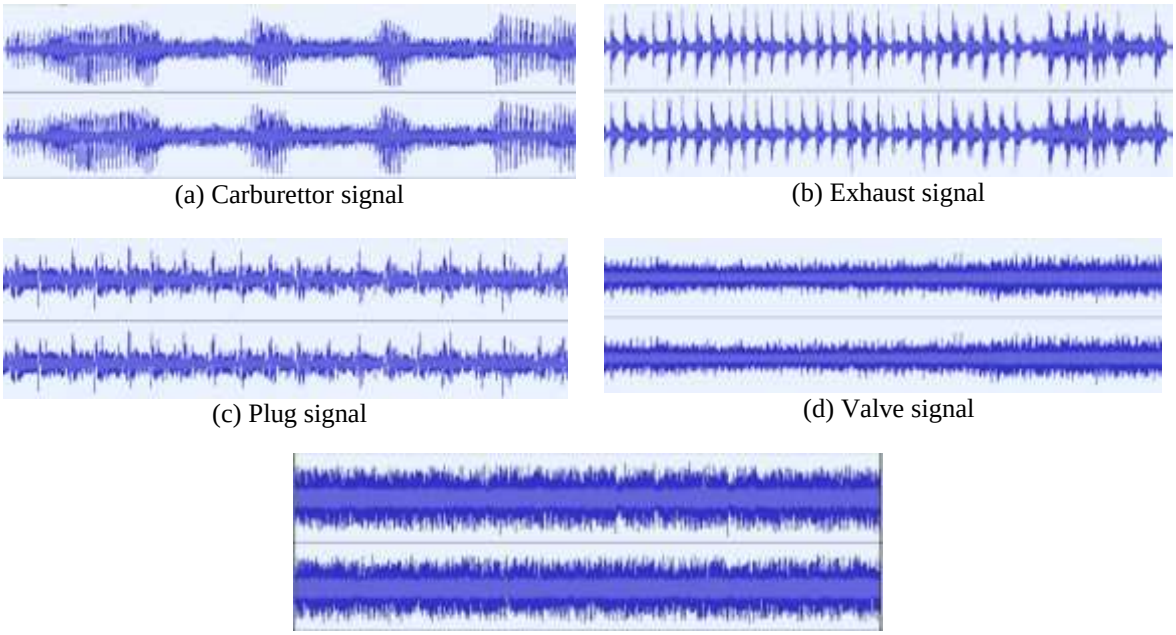
4. RESULTS AND DISCUSSION

In this section, the results obtained at various stages of system development are presented. The recorded audio signals for various faults and normal operating conditions are presented. The transformed signals used as input for the deep learning models are also shown. Additionally, the performance evaluation results for accuracy, precision, recall, and F1-score are presented for each model-spectrogram pair. The performance evaluation results of the MobileNet and AlexNet deep learning models are discussed, using accuracy, precision, recall, and F1-Score, as well as the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metrics adapted for multiclass cases. The inclusion and discussion of the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were adapted for multiclass scenarios to extend model evaluation beyond accuracy.

4.1 Waveforms of Generator Audio Signals and Spectrograms

The signal waveforms of the recorded fault audio signals and the respective spectrograms used for fault type identification and data pre-processing for model development were obtained.

The signal waveforms showing temporal characteristics of the fault signals, variations in amplitude, and frequency patterns associated with different fault conditions are shown in Figure 5. The spectral signatures of the audio signals are presented in Figure 6.

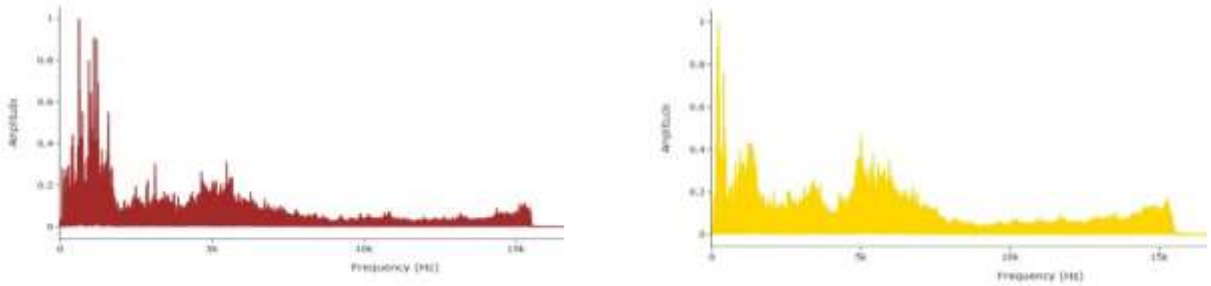


(a) Carburettor signal (b) Exhaust signal

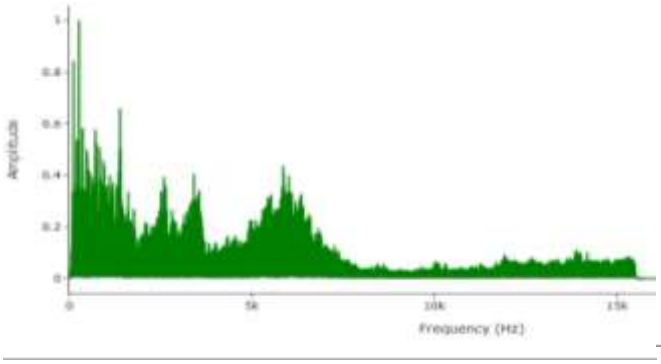
(c) Plug signal (d) Valve signal

(e) Normal signal

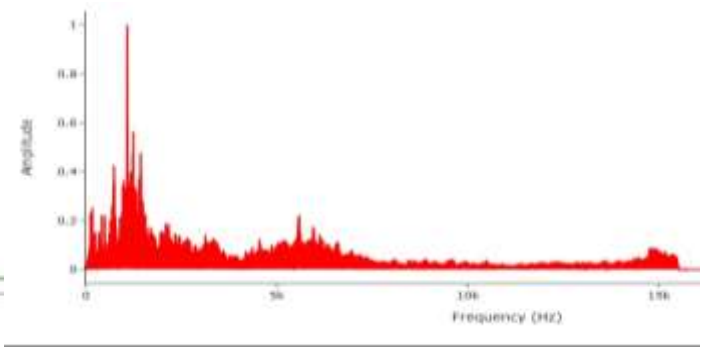
Figure 5: Waveforms of recorded generator audio fault signals



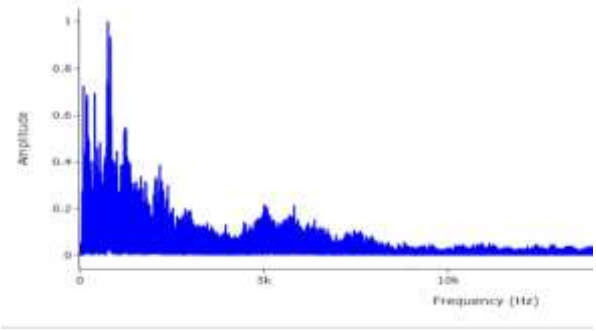
(a) Carburetor signal (b) Exhaust signal



(c) Plug signal



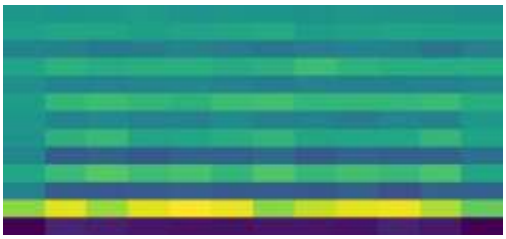
(d) Valve signal



(e) Normal signal

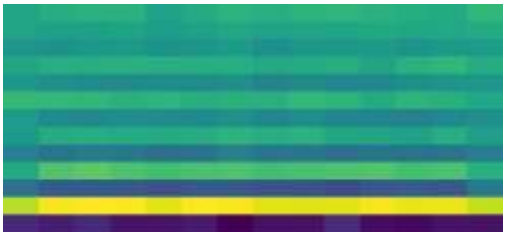
Figure 6: Spectral signatures of recorded generator audio fault signal

Figures 7, 8, and 9 show the respective MFCC, CWT, and STFT spectrograms generated for the recorded generator fault signals and used for training deep learning models.

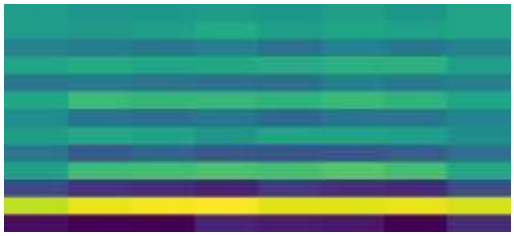


(a) Carburetor signal

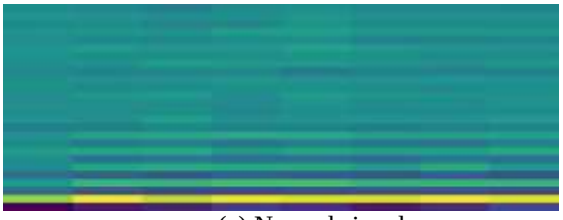
(b) Exhaust signal



(c) Plug signal

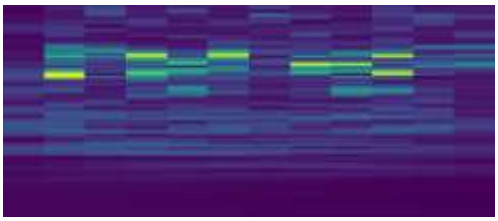


(d) Valve signal

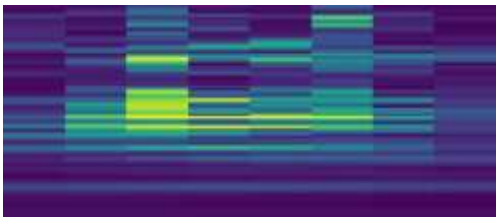


(e) Normal signal

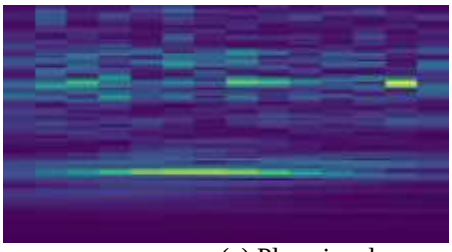
Figure 7: MFCC spectrogram for alexnet and mobilenet deep learning models



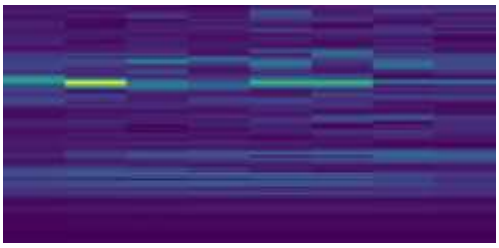
(a) Carburetor signal



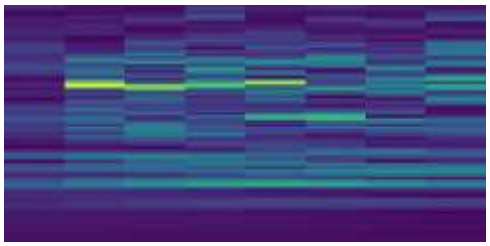
(b) Exhaust signal



(c) Plug signal

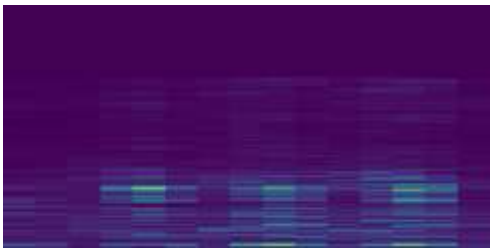


(d) Valve signal

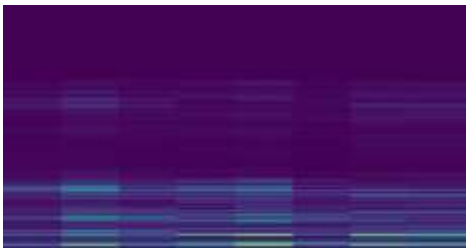


(e) Normal signal

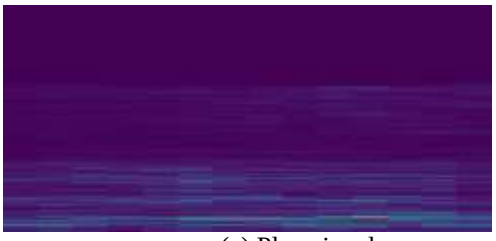
Figure 8: CWT spectrogram for AlexNet and MobileNet deep learning models



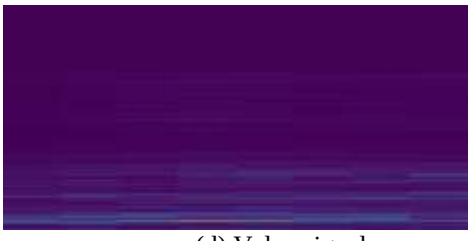
(a) Carburetor signal



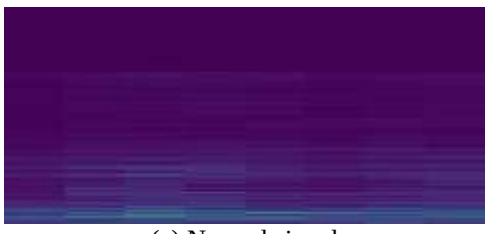
(b) Exhaust signal



(c) Plug signal



(d) Valve signal



(e) Normal signal

Figure 9: STFT spectrogram for AlexNet and MobileNet deep learning models

4.2 Performance Evaluation of Deep Learning Models

The performances of the AlexNet and MobileNet deep learning models, developed for all 2D spectrogram images generated from the validation test dataset, were evaluated using accuracy, precision, F1-score, and ROC/AUC evaluation metrics. The performance evaluation was done on the three spectrogram transformations: Mel-Frequency Cepstral Coefficients (MFCC), Continuous Wavelet Transform (CWT), and Short-Time Fourier Transform (STFT).

4.2.1 Performance metrics of AlexNet and MobileNet deep learning models using MFCC spectrograms

Table 1 summarizes the results obtained from the accuracy, precision, F1-score, and AUC evaluation metrics for the performance of the AlexNet and MobileNet deep learning models when processing Mel-frequency cepstral coefficient (MFCC) spectrograms. Figure 10 shows the Receiver Operating Characteristic (ROC) trajectories for the AlexNet and MobileNet models, accompanied by their respective Area Under the Curve (AUC) quantifications, which show the trade-offs between true positive rates and false positive rates across the classification thresholds.

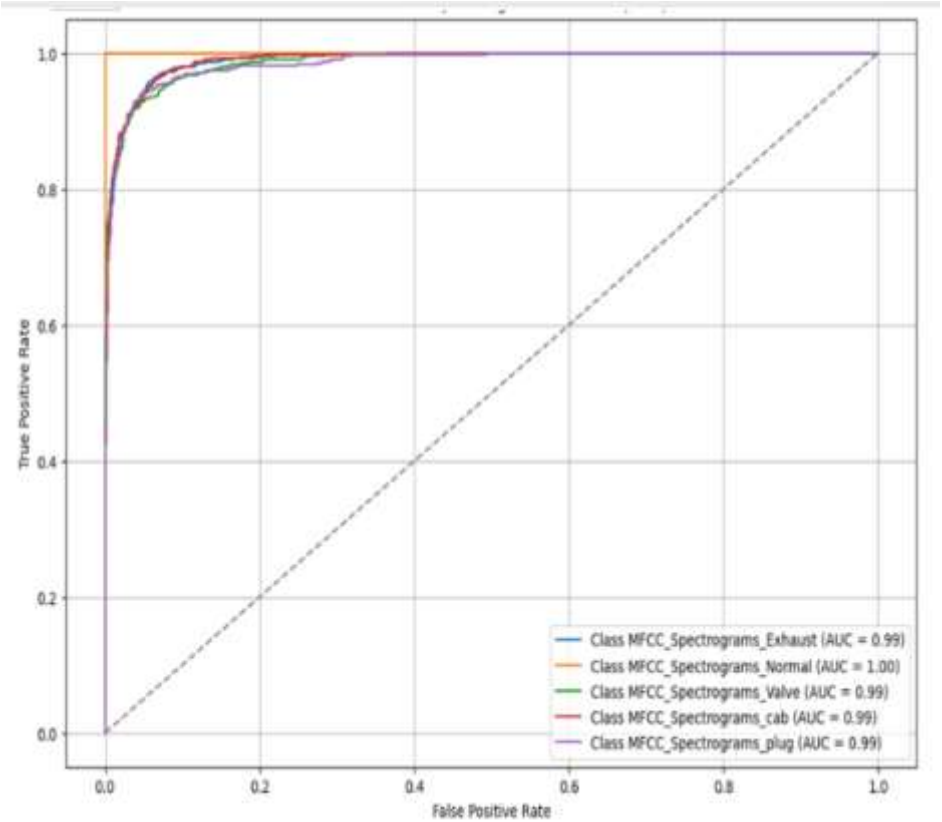
Table 1: Performance metrics for AlexNet and MobileNet deep learning models with MFCC spectrograms

Model	Accuracy	Precision	Recall	F1-Score	AUC value
MobileNet	90%	90%	90%	90%	99%
AlexNet	59%	62%	59%	58%	85.6%

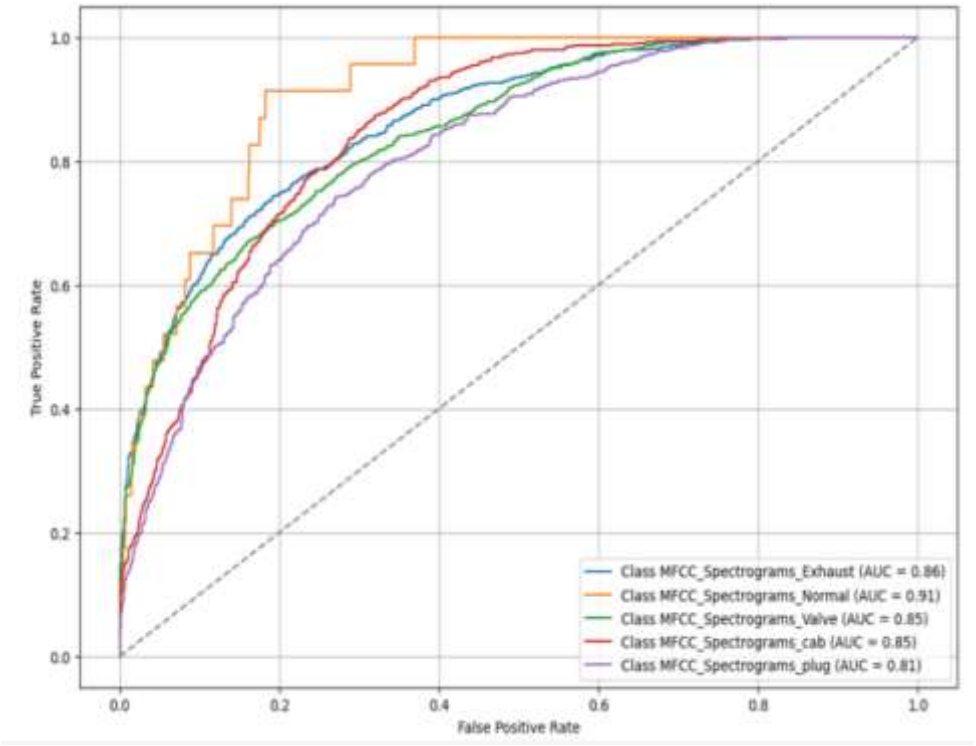
Figures 10, 11, and 12 show the plot of the ROC curve. The X-axis represents the False Positive Rate (FPR), which is the proportion of negative samples (normal cases) incorrectly classified as positive (faults); the Y-axis represents the True Positive Rate (TPR/Sensitivity) - the proportion of actual fault cases (positive samples) correctly identified; and the diagonal line represents a random guess (AUC = 0.5). Each curve corresponds to a different fault class, as shown by the legend, with its associated AUC (Area Under the Curve) value. The AUC represents the ability of the model to distinguish between classes. AUC ranges from 0 to 1, where 1.0 = perfect classification, 0.5 = no discrimination (random guess), and < 0.5 = worse than random.

4.2.2 AUC-ROC curve for MobileNet and AlexNet deep learning models with MFCC spectrogram

The plot in Figure 10(a) displays the ROC curves and AUC values for a MobileNet classification model using MFCC spectrograms as input features. It is observed that the ROC curves for all classes are close to the top-left corner. All the curves are well above the dashed diagonal line representing a random classifier (AUC = 0.5). This indicates very high true positive rates and low false positive rates across all thresholds, indicating that the model exhibits strong performance. The AUC scores for MobileNet using MFCC Spectrograms from the ROC plot in Figure 10(a) are summarized in Table 2



(a) MobileNet Model



(b) AlexNet Model

Figure 10: ROC curve for AlexNet and MobileNet deep learning models with MFCC Spectrogram

Table 2: ROC-AUC scores for MobileNet using MFCC Spectrograms

MobileNet Class	AUC Score
MFCC_Spectrograms_Exhaust	0.99
MFCC_Spectrograms_Normal	1.00
MFCC_Spectrograms_Valve	0.99
MFCC_Spectrograms_Caburettor	0.99
MFCC_Spectrograms_Plug	0.99

The MobileNet model has an AUC of approximately 1.0 for all fault cases. This means that the model can almost perfectly distinguish between each class. For the Normal class, MobileNet achieves an AUC of 1.00, indicating perfect separability between the "Normal" class and all other classes. This is because the ROC curve for this class passes through the top-left corner (0,1), and the model makes no false positives or false negatives at the decision threshold. It also indicates that a threshold exists for the model's prediction scores, where neither false positives nor false negatives occur for the "Normal" class. The model's prediction confidence for the Normal class is strong and well-separated from that of the other classes. This is important because it suggests MobileNet has learned the unique and easily distinguishable features of the Normal class in the MFCC spectrograms. This implies that the MobileNet model's high AUC values across all classes reveal well-represented features by MFCC, effective learning of class-specific patterns by MobileNet, and minimal overfitting/underfitting based on ROC performance alone.

In Figure 10(b), the ROC curve image illustrates the performance of the AlexNet model on a multiclass classification task, utilizing MFCC spectrograms as input features. The ROC-AUC Plot displays the class-wise performance for five different sound categories in the generator fault detection and classification task. Table 3 presents a summary of AUC scores from the ROC plot in Figure 10(b) for AlexNet using MFCC Spectrograms.

The Normal condition with AUC = 0.91 is the best-performing class in this model. This means that the model is relatively strong at distinguishing "Normal" spectrograms from others. High separability indicates the ability to detect healthy or baseline conditions effectively. The Exhaust fault condition with AUC = 0.86 gives a reasonably good class separation. The model can effectively detect exhaust-related patterns, but there may still be some overlap with other class features. The Valve & Caburettor fault condition with AUC = 0.85 indicates moderate performance, which is acceptable, but the model may confuse some Valve or Caburettor samples with others.

Table 3: ROC-AUC scores for AlexNet using MFCC Spectograms

AlexNet Class	AUC Score
MFCC_Spectrograms_Exhaust	0.56
MFCC_Spectrograms_Normal	0.91
MFCC_Spectrograms_Valve	0.85
MFCC_Spectrograms_Caburettor	0.85
MFCC_Spectrograms_Plug	0.81

The Plug fault with AUC = 0.81 has the weakest class in terms of separability. This means that the Plug class is more challenging to distinguish due to overlapping features in the spectrogram. Even though all ROC curves lie well above the diagonal, indicating the model performs better than random guessing, AlexNet's performance is weaker compared to the MobileNet model, which has AUCs close to 0.99 or 1.00. This suggests less effective feature learning or generalization, and possibly indicates that AlexNet is less suited to this task. A step-by-step comparison of the performance of MobileNet and AlexNet on the MFCC spectrogram classification task using ROC-AUC is shown in Table 4.

Table 4: ROC-AUC comparison table for Mobilenet and AlexNet on the MFCC spectrogram

Class	MobileNet AUC	AlexNet AUC	Difference (MobileNet – AlexNet)
Exhaust	0.99	0.86	+0.13
Normal	1.00	0.91	+0.09
Valve	0.99	0.85	+0.14
Caburettor	0.99	0.85	+0.14
Plug	0.99	0.81	+0.18

According to the values presented in Table 4, MobileNet outperforms AlexNet significantly across all classes, with differences ranging from +0.09 to +0.18 in AUC. The most significant performance gap is for the Plug class (+0.18), which was the weakest among AlexNet's performance gaps. This indicates MobileNet is much better at learning subtle distinctions in that class. Even the Normal class, where AlexNet performs best, is outperformed by MobileNet.

4.2.3 Performance metrics of AlexNet and MobileNet deep learning models using CWT spectrograms

The summary of the results obtained from the accuracy, precision, F1-score, and AUC evaluation metrics for the performance of the AlexNet and MobileNet deep learning models trained on Continuous Wavelet Transform (CWT) time-frequency representations is shown in Table 5.

Table 5: Performance metrics for AlexNet and MobileNet deep learning models with CWT Spectrograms

Model	Accuracy	Precision	Recall	F1-Score	AUC value
MobileNet	86%	87%	86%	86%	98%
AlexNet	55%	59%	55%	52%	84.8%

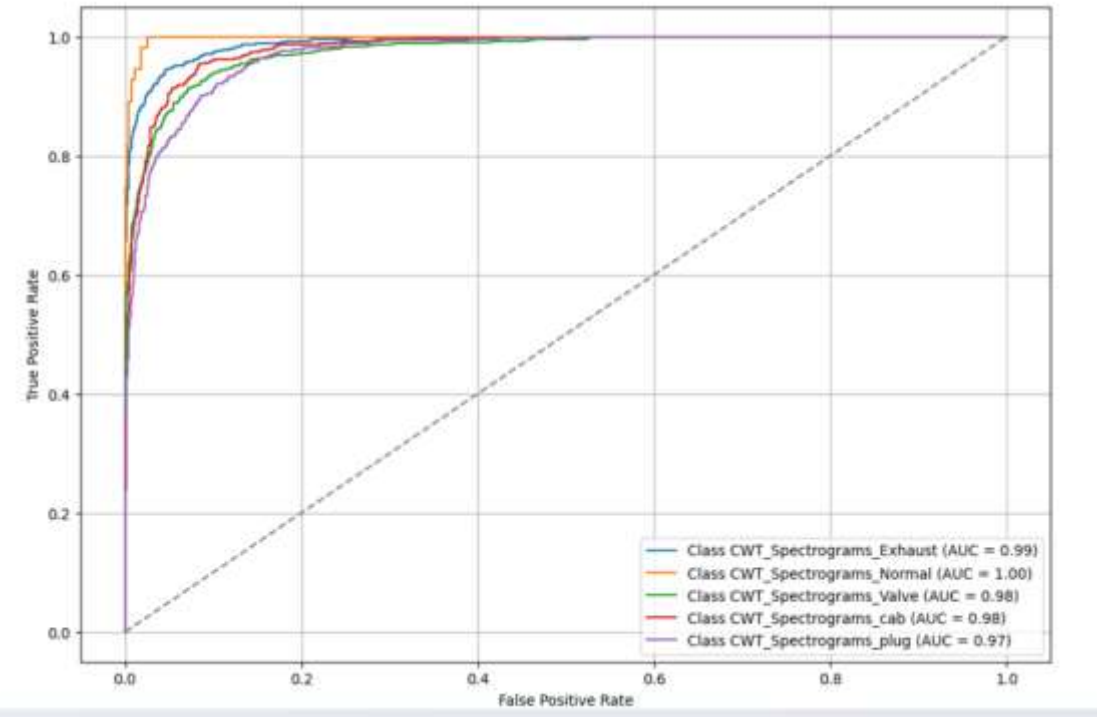
Table 6: ROC-AUC scores for MobileNet using CWT Spectograms

MobileNet Fault Class	AUC Score
CWT_Spectrograms_Exhaust	0.99
CWT_Spectrograms_Normal	1.00
CWT_Spectrograms_Valve	0.98
CWT_Spectrograms_Caburettor	0.98
CWT_Spectrograms_Plug	0.97

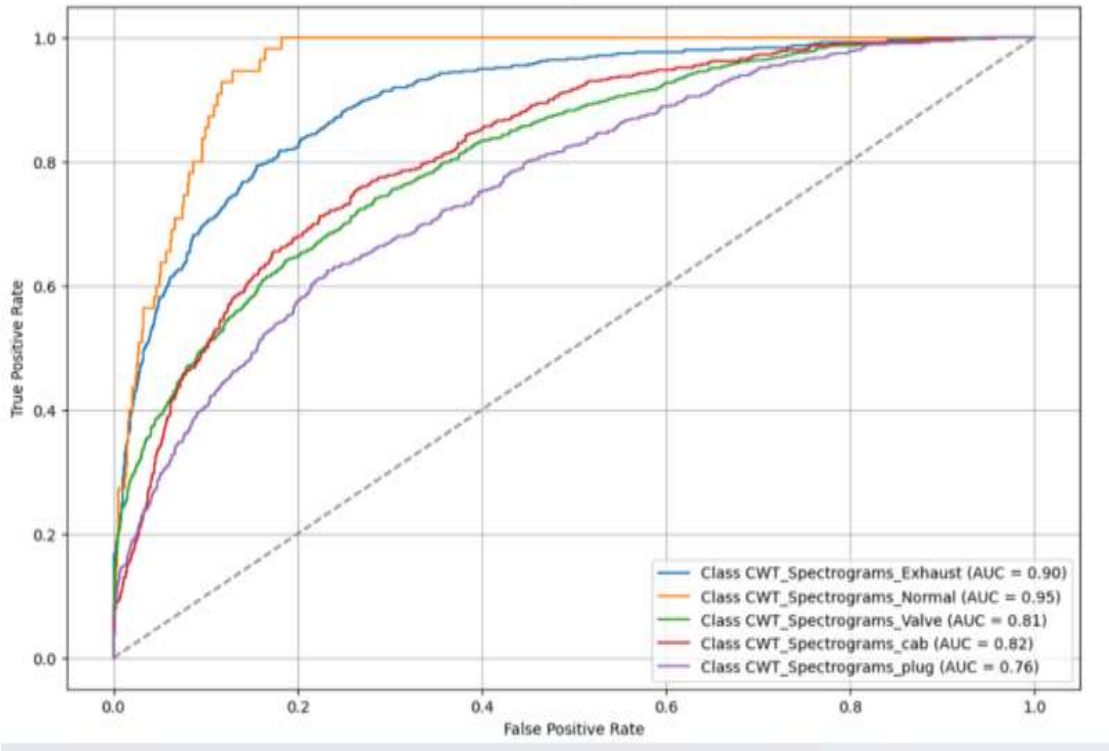
4.2.4 ROC-AUC curve for AlexNet and MobileNet deep learning models with CWT spectrograms

Figure 11 displays the Receiver Operating Characteristic (ROC) trajectories for the models, along with the Area Under the Curve (AUC) quantifications that reflect the discriminative capacity of AlexNet and MobileNet Deep Learning Models with CWT Spectrograms across different classification thresholds.

The score for the ROC Curve for MobileNet and AlexNet Deep Learning Models with CWT Spectrograms as input features, shown in Figure 11(a) and 11(b) respectively, is summarized in Tables 6 and 7, respectively. Table 5 presents the compilation of AUC scores for MobileNet using CWT spectrograms extracted from the ROC plot in Figure 11a.



(a) MobileNet Model



(b) AlexNet Model

Figure 11: ROC curve for AlexNet and MobileNet deep learning models with CWT spectrograms

In the class-wise performance, the Normal AUC of 1.00 indicates that MobileNet accurately distinguished the "Normal" class from the CWT spectrogram without error. The Exhaust (0.99), Valve (0.98), Carburetor (0.98), and Plug (0.97) have exceptionally high AUC values, indicating that the model exhibits a very strong discriminative ability across all fault types. The shape of the ROC Curves is at the top-left corner, indicating that the True Positive Rate (TPR) is high, the False Positive Rate (FPR) is low, and the trade-off between sensitivity and specificity is Minimal.

Table 7: ROC-AUC scores for AlexNet using CWT Spectograms

AlexNet Fault Class	AUC Score
CWT_Spectrograms_Exhaust	0.90
CWT_Spectrograms_Normal	0.95
CWT_Spectrograms_Valve	0.81
CWT_Spectrograms_Caburettor	0.82
CWT_Spectrograms_Plug	0.76

The values in Table 7 indicate that the AlexNet model, utilizing CWT spectrograms as input, exhibits a strong classification capability overall, particularly for normal operation and exhaust faults, suggesting that deviations from normal conditions are generally well captured. The analysis of the model's performance shows the Normal Class with AUC = 0.95. The model performs very well in distinguishing normal signals from fault signals. This high AUC implies strong classification confidence and a low false positive rate. The Exhaust Class with AUC = 0.90 indicates that the model also shows excellent performance in detecting exhaust faults. This suggests that exhaust faults generate distinctive spectrogram features.

However, it has more difficulty distinguishing between plug faults, Valve, and carburetor faults, as shown by their relatively lower AUC values for the carburetor and Valve classes, with AUC ≈ values of approximately 0.81–0.82 indicating moderate to good classification ability. This suggests overlapping signal patterns or that the features from the CWT spectrograms are less distinctive. The Class "Plug" with the AUC = 0.76 is the lowest-performing class. It likely indicates that spark plug-related faults are harder to detect using the current CWT-AlexNet approach.

Table 8: ROC-AUC comparison table for MobileNet and AlexNet on the CWT spectrogram

Class	MobileNet AUC	AlexNet AUC	Difference (MobileNet – AlexNet)
Exhaust	0.99	0.90	+0.09
Normal	1.00	0.95	+0.05
Valve	0.98	0.81	+0.17
Caburettor	0.98	0.82	+0.16
Plug	0.97	0.76	+0.21

According to the comparison in Table 8, MobileNet significantly outperforms AlexNet across all classes, with differences ranging from +0.09 to +0.21 in ROC-AUC. The largest performance gap is for the Plug class (+0.21), which was the weakest for AlexNet. This indicates MobileNet is much better at learning subtle distinctions in that class. Even the Normal class, where AlexNet performs best, is outperformed by MobileNet.

4.2.5 Performance metrics of AlexNet and MobileNet deep learning models using STFT spectrograms

Table 9 shows the performance metrics—classification accuracy, precision (positive predictive value), recall (sensitivity), and F1-score (harmonic precision-recall equilibrium)—for the AlexNet and MobileNet deep learning architectures trained on Short-Time Fourier Transform (STFT) spectrograms.

Table 9: Performance metrics for AlexNet and MobileNet deep learning models with STFT spectrograms

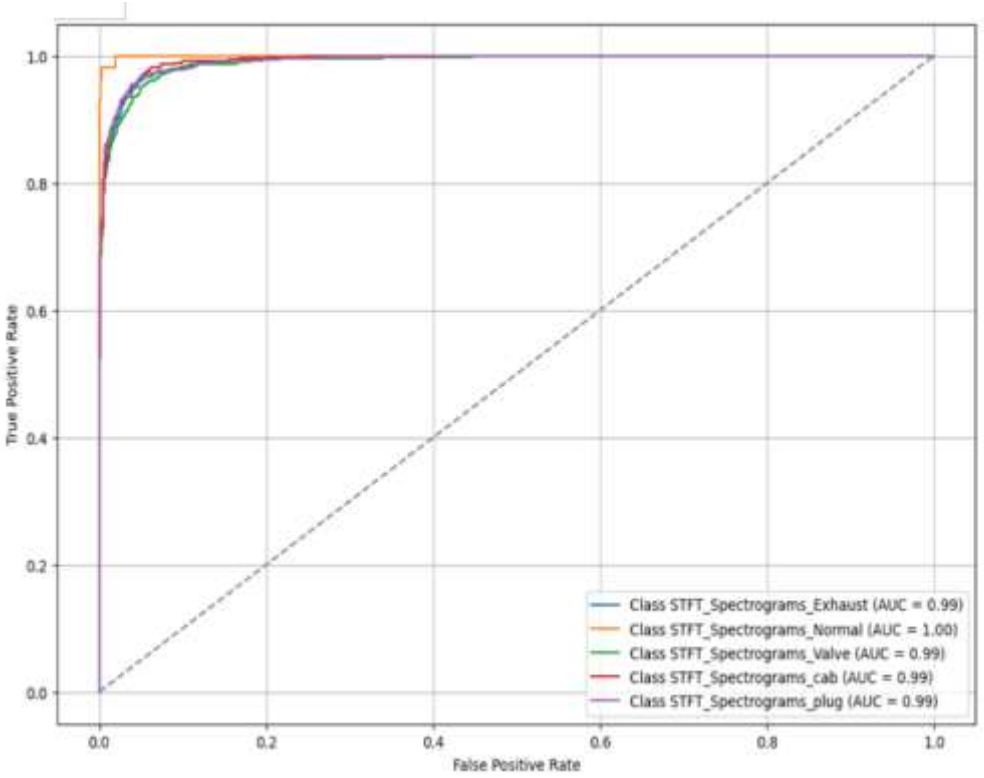
Model	Accuracy	Precision	Recall	F1-Score	AUC value
MobileNet	92%	92%	92%	92%	99%
AlexNet	56%	625	56%	55%	88.4%

4.2.6 ROC-AUC curve for MobileNet and AlexNet deep learning Models with STFT spectrogram

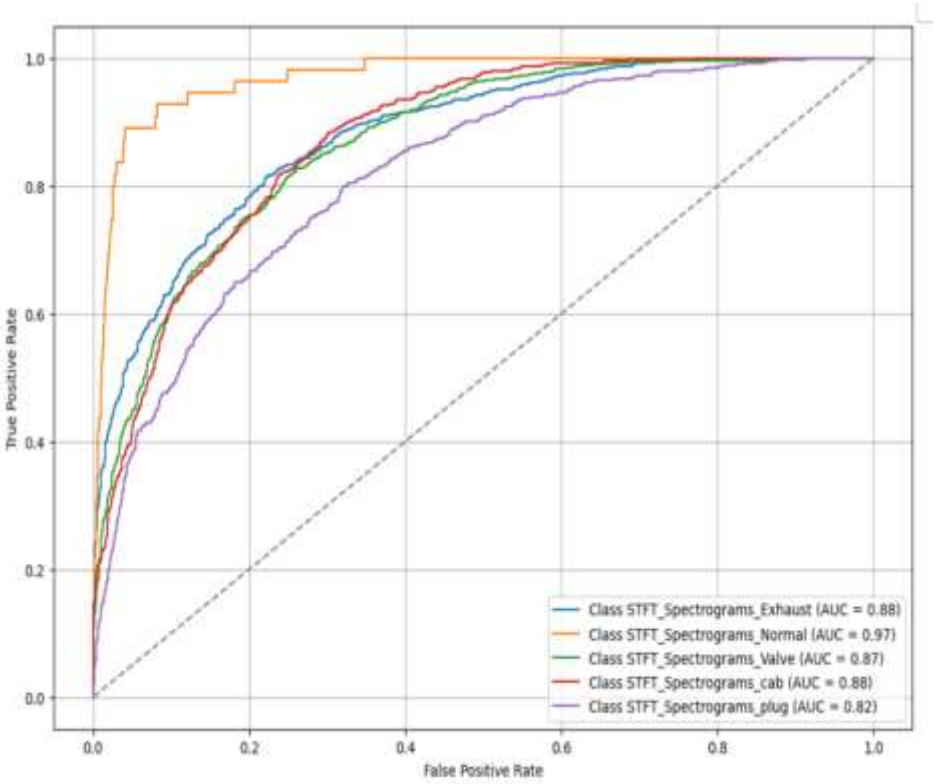
Figure 12 shows the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) metric plots that quantify AlexNet and MobileNet Deep Learning Models' ability to distinguish between classes under different decision thresholds. Together, these diagnostics assess robustness in scenarios requiring temporal-spectral feature discrimination by making trade-offs between false positives and true positives.

The score for the ROC Curve for MobileNet and AlexNet Deep Learning Models with STFT spectrograms as input features, shown in Figure 12, is summarized in Tables 10 and 11, respectively. Table 10 is the compilation of the ROC-AUC scores for MobileNet using the STFT spectrograms. The Key Observations from the results presented in Table 10 are:

i. The MobileNet on the STFT spectrogram shows outstanding performance across all classes. The AUC scores are either 0.99 or 1.00, indicating that the MobileNet model with STFT spectrograms achieves near-perfect classification.
ii. The model perfectly distinguished the normal operations (AUC = 1.00) from all fault types—this is highly desirable in fault diagnosis systems to avoid false alarms.



(a) MobileNet Model



(b) AlexNet Model

Figure 12: ROC curve for AlexNet and MobileNet deep learning models with STFT spectrograms.

iii. The tight clustering of all curves near the top-left shows the model maintains high accuracy with very low false positive and false negative rates / minimal class Confusion.
iv. STFT spectrograms appear to provide highly discriminative time-frequency features for MobileNet with robust generalization. Likely due to the local time-frequency precision of STFT, combined with MobileNet's ability to capture spatial features.

Table 10: ROC-AUC scores for MobileNet on the STFT spectrogram

MobileNet Class	AUC Score
STFT_Spectrograms_Exhaust	0.99
STFT_Spectrograms_Normal	1.00
STFT_Spectrograms_Valve	0.99
STFT_Spectrograms_Caburettor	0.99
STFT_Spectrograms_Plug	0.99

The ROC curve from Figure 12(b) shows the performance of the AlexNet deep learning model using Short-Time Fourier Transform (STFT) spectrograms as input features for fault diagnosis in the 5kVA power generating sets. The values of the ROC-AUC curve for the AlexNet deep learning model are presented in Table 11.

Table 11: ROC-AUC scores for AlexNet on the STFT spectrogram

AlexNet Class	AUC Score
STFT_Spectrograms_Exhaust	0.88
STFT_Spectrograms_Normal	0.97
STFT_Spectrograms_Valve	0.87
STFT_Spectrograms_Caburettor	0.88
STFT_Spectrograms_Plug	0.82

For the insights and analysis derived from the ROC-AUC values for AlexNet on the STFT spectrogram, the Class: Normal, represented by the Orange curve with an AUC of 0.97, demonstrated exceptional performance by the AlexNet model. It also indicates a Low false positive rate and high sensitivity. For the Exhaust, Valve, and Carburettor classes (AUC ~ 0.87–0.88), the model demonstrated good performance, indicating that the STFT-based spectrograms provide sufficient discriminative features, although they are slightly less accurate than the Normal class. The Class "Plug" with the Purple curve, AUC = 0.82, was the least well classified among the group. This could be due to overlapping patterns or subtler features, making plug faults harder to detect with AlexNet using STFT.

Table 12 presents a comparison of MobileNet and AlexNet, using STFT spectrograms as input features for fault diagnosis in 5kVA power generating sets.

Table 12: Comparison of MobileNet and AlexNet using STFT spectrograms

Fault Class	MobileNet AUC	AlexNet AUC	Difference (MobileNet – AlexNet)
Exhaust	0.99	0.88	+0.11
Normal	1.00	0.97	+0.03
Valve	0.99	0.87	+0.12
Caburettor	0.99	0.88	+0.11
Plug	0.99	0.82	+0.17

The comparison results in Table 11 show that MobileNet significantly outperforms AlexNet across all classes using STFT features. The AUC values are extremely high (≥ 0.99), indicating near-perfect classification, and are especially important for hard-to-detect faults like Plug (AUC: 0.99 vs. 0.82). The Normal class achieves perfect classification (AUC = 1.00) with MobileNet. This is critical for avoiding false alarms and ensuring reliable system health assessments. AlexNet still shows good performance (AUC > 0.85 in most classes), but lags behind in fine-grained discrimination, particularly in the Plug class.

4.2.7 Comparison of AUC-ROC curve for MobileNet and AlexNet deep learning models with MFCC, CWT, and STFT spectrograms

The comparison of MobileNet's performance across three different audio feature extraction techniques: MFCC, CWT, and STFT spectrograms for fault diagnosis in 5kVA power generating sets is shown in Table 13.

Table 13: The comparison of MobileNet's performance across the MFCC, CWT, and STFT spectrograms

Fault Class	MFCC AUC	CWT AUC	STFT AUC	Best Feature Type
Exhaust	0.86	0.90	0.99	STFT
Normal	0.91	0.95	1.00	STFT
Valve	0.85	0.81	0.99	STFT
Caburettor	0.85	0.82	0.99	STFT
Plug	0.81	0.76	0.99	STFT

The comparison of MobileNet's performance with the MFCC, CWT, and STFT input spectrograms, as presented in Table 13, shows that the STFT Spectrograms outperform MFCC and CWT across all fault classes when used with MobileNet. This indicates that the STFT spectrogram captured finer time-frequency representations and so is better suited for MobileNet's convolutional layers. Though CWT's performance was consistently inferior to STFT, it was slightly better than MFCC in the Exhaust and Normal classes. This indicates that CWT could capture more detailed wavelet-based features, but it may be less compatible with MobileNet's structure. With respect to the Plug and Valve classes, though they were typically harder to classify, they showed significant improvement with STFT, with the AUC moving from the range ~0.76–0.85 to 0.99. A similar comparison across the MFCC, CWT, and STFT input spectrograms is done for AlexNet and presented in Table 14.

Table 14: The comparison of AlexNet's performance across the MFCC, CWT, and STFT spectrograms

Fault Class	MFCC AUC	CWT AUC	STFT AUC	Best Feature Type
Exhaust	0.86	0.90	0.88	CWT
Normal	0.91	0.95	0.97	STFT
Valve	0.85	0.81	0.87	STFT
Caburettor	0.85	0.82	0.88	STFT
Plug	0.81	0.76	0.82	STFT

The observations made from the ROC-AUC values obtained across MFCC, CWT, and STFT input spectrograms in relation to the performance of the AlexNet model are as follows: -

i. The AUC values with STFT range from 0.82 to 0.97, showing more consistent results across all fault types.
ii. AlexNet's performance in identifying the Normal, Valve, Cab, and Plug classes was highest with the STFT input Spectrograms overall.
iii. From the AUC value of Exhaust (0.90) for the CWT input spectrogram, which is slightly higher than STFT (0.88) and MFCC (0.86), it suggests that the wavelet-based representations might capture useful transient signals for the Exhaust class.
iv. The MFCC input spectrograms presented a weaker overall performance for AlexNet, most especially in faults like Plug (0.81) and Valve (0.85), which appeared harder to detect. It suggests that this spectrogram may lack the rich time-frequency resolution needed for AlexNet's convolutional architecture.

4.2.8 Comparative Accuracy Performance of MobileNet and AlexNet Deep Learning Models

A side-by-side comparison of the accuracy scores for MobileNet and AlexNet deep learning models, as shown in Figure 13, is discussed. The performance evaluation of the two deep learning architectures on the MFCC, CWT, and STFT spectrogram transformations is indicated by the legends: blue for MFCC, red for CWT, and green for STFT.

The bar chart in Figure 13 compares the accuracy (%) of MobileNet and AlexNet deep learning models using MFCC (Mel Frequency Cepstral Coefficients), CWT (Continuous Wavelet Transform), and STFT (Short-Time Fourier Transform) as input spectrogram features. It was observed that MobileNet's Performance was MFCC: ~90% accuracy, CWT: ~86% accuracy, and STFT: ~92% accuracy. This indicates that MobileNet performed very well across all three feature extraction methods, with STFT yielding the highest accuracy, closely followed by MFCC. CWT shows slightly lower performance but is still above 85%. The performance of the AlexNet models showed MFCC with ~59% accuracy, CWT with ~54% accuracy, and STFT with ~56% accuracy, indicating that AlexNet's performance was significantly lower across all features compared to MobileNet. Among the three spectrograms, MFCC yielded the best result for AlexNet; however, it still fell below 60%, indicating poor classification performance. The above accuracy scores are shown in Table 15. The accuracy results obtained suggest that MobileNet performs better across all feature types, making it better suited for this specific

classification task. This may be due to its more optimized and lightweight architecture, which may generalize better on spectrogram data.

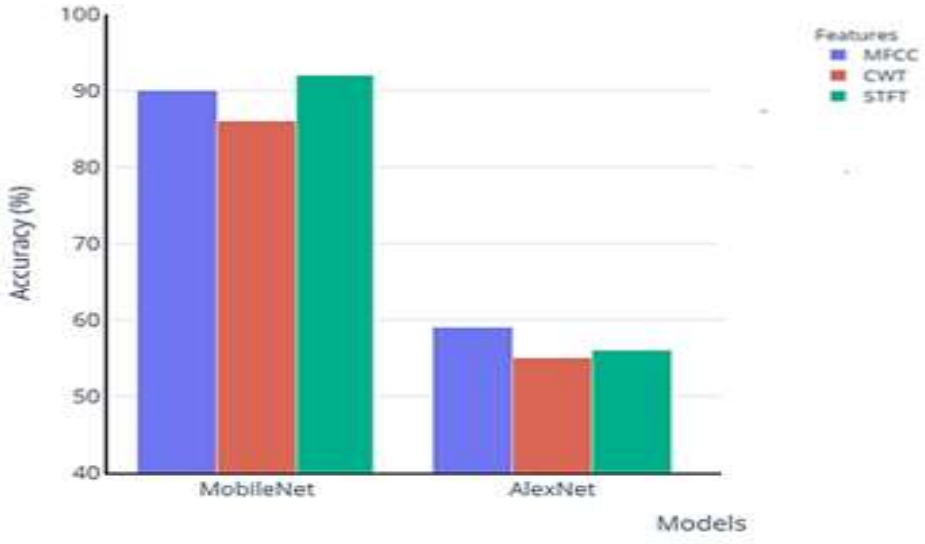


Figure 13: A Plot comparing the accuracy performance of the MobileNet and AlexNet deep learning models

Table 15: Comparative analysis of the developed models

Model	Accuracy
MobileNet_MFCC	90%
AlexNet_MFCC	59%
MobileNet_CWT	86%
AlexNet_CWT	55%
MobileNet_STFT	92%
AlexNet_STFT	56%

Table 16 presents a comparative analysis of this work with other studies in the field of fault diagnosis using machine learning. To the best of our knowledge, no research has been carried out on the category of DaSSC-PGS considered in this work. Thus, the comparative analysis is conducted on a broad scale to showcase the benefits of using machine learning for fault diagnosis across all categories of machinery.

Table 16: Comparative analysis of the best-performing model with models from other works

Author	Model	Accuracy	Application
This work	MobileNet_STFT	92%	Domestic Generator sets (5kVA)
[18]	LSTM	99.65%	Electrical Power Transmission System
[4]	Deep CNN	99.07%	Lab-scale gas turbine
[14]	Generative adversarial networks (GANs)	98.9%	Power Generation Plant
[3]	CNN, vision transformer (ViT) model	CNN < 70%, ViT model over 90%.	Electric power system
[15]	SVM	95%	Industrial Refrigeration Systems
[16]	Optimised SSAE model	99%	wind turbine gearbox

From the results shown in Table 15, the MobileNet_STFT model developed in this work achieved an accuracy of 92% for fault diagnosis in domestic generator sets (5kVA). Other models in the literature reported higher accuracies, such as - LSTM (99.65%) for electrical power transmission systems [18], Deep CNN (99.07%) for lab-scale gas turbines [4], GANs (98.9%) for power generation plants [14], Optimised SSAE (99%) for wind turbine gearboxes [16], and SVM (95%) for

industrial refrigeration systems [15], and [3] showed that CNNs performed poorly (<70%), while Vision Transformers (ViT) achieved over 90% accuracy on electric power systems.

While the MobileNet_STFT model does not outperform all other models in terms of raw accuracy, it is competitive and tailored to a specific application domain (small domestic generators), which differs from the typically larger-scale or industrial systems studied in other works. The slightly lower accuracy may reflect the challenges unique to this domain, such as noisier signals or limited data, while still offering a lightweight and efficient solution. On the other hand, considering that most of the accuracies are above 90%, these results show great promise in the advantage of using machine learning for fault diagnosis and detection.

5. CONCLUSIONS

This study compared the performance of the MobileNet and AlexNet deep learning models for fault diagnosis in twenty-five 5kVA power generating sets using three different time-frequency spectrogram features. The approach used involved recording one normal and four faulty conditions of the generator sets over a 1-minute and 30-second period. The audio signals were pre-processed and transformed into MFCC, CWT, and STFT spectrograms, which were then fed into the models for feature extraction and training. The models were subsequently tested on the portion of the data not used for training. The performance of the models was evaluated using F1-score, accuracy, recall, precision, and ROC-AUC metrics.

The results show that MobileNet, when paired with STFT spectrograms, achieves the highest classification performance, reaching 92% accuracy and AUC scores of 0.99 or higher across all fault classes. Among all the models evaluated in this study, MobileNet demonstrated the highest reliability and efficiency for real-time, edge-based fault detection in domestic generator applications. In contrast, AlexNet consistently underperformed, achieving a maximum accuracy of 59% with MFCC spectrograms. While it showed relatively better class-wise discrimination for Normal and Exhaust faults using CWT spectrograms, its ability to identify subtler faults such as Valve, Carburetor, and Plug was limited. This poor performance is likely due to overlapping spectral features, as indicated by lower AUC values. Overall, the findings strongly support the adoption of a MobileNet + STFT framework for developing practical, lightweight, and high-performing generator fault diagnosis systems—particularly suited to the needs of the domestic energy landscape.

6. RECOMMENDATIONS

To further improve classification performance, most especially for the Plug faults that were difficult to detect, future systems may benefit from a hybrid feature fusion approach that combines multiple spectrogram features, including CWT, MFCC, and STFT. Such fusion has the potential to capture complementary acoustic characteristics and strengthen the separability of fault classes with subtle signal overlaps. Although MobileNet achieved superior performance in this work, exploring other deep learning architectures, such as ResNet, DenseNet, and EfficientNet, may also offer improved accuracy and robustness, particularly in handling complex acoustic patterns. They may also be beneficial in enhancing training data for underperforming classes.

The dataset used in this work was limited to 5kVA DaSSC-PGS. Expanding future datasets to include a wider variety of generator capacities, other categories of small-scale and industrial power systems, and diverse operating environments will be crucial for improving model generalization and applicability across various use cases in both domestic and commercial contexts. This may contribute to the advancement of standardized, intelligent, and scalable audio-based fault diagnosis systems, ultimately supporting more reliable and sustainable energy solutions. Using real-time deployment to investigate latency, computational efficiency, and energy consumption on mobile or embedded platforms is crucial for establishing the practicality of lightweight CNNs for everyday use in resource-constrained environments. Insights into possible performance trade-offs and deployment feasibility can be explored through further comparative analysis involving recurrent neural networks or hybrid deep learning models, alongside traditional machine learning techniques.

REFERENCES

[1] Shang, H., Liu, Z., Wei, Y., & Zhang, S. (2024). A novel fault diagnosis method for a power transformer based on multi-scale approximate entropy and optimized convolutional networks. Entropy, 26(3), 186

[2] Henriquez, P., Alonso, J. B., Ferrer, M. A., & Travieso, C. M. (2013). Review of automatic fault diagnosis systems using audio and vibration signals. IEEE Transactions on systems, man, and cybernetics: Systems, 44(5), 642-652.

[3] Yoon, D. H., & Yoon, J. (2024). Development of a real-time fault detection method for electric power system via transformer-based deep learning model. International Journal of Electrical Power & Energy Systems, 159, 110069.

[4] Nashed, M. S., Renno, J., & Mohamed, M. S. (2024). Fault classification using convolutional neural networks and color channels for time-frequency analysis of acoustic emissions. Journal of Vibration and Control, 30(9-10), 2283-2300.

[5] Kowalski, J., Krawczyk, B., & Woźniak, M. (2017). Fault diagnosis of marine 4-stroke diesel engines using a one-vs-one extreme learning ensemble. Engineering Applications of Artificial Intelligence, 57, 134-141.

[6] Mitiche, I., Nesbitt, A., Conner, S., Boreham, P., & Morison, G. (2020). 1D-CNN based real-time fault detection system for power asset diagnostics. IET Generation, Transmission & Distribution, 14(24), 5766-5773.

[7] Vilela, R., Metrolho, J. C., & Cardoso, J. C. (2004). Machine and industrial monitorization system by analysis of acoustic signatures. In Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference (IEEE Cat. No. 04CH37521) (Vol. 1, pp. 277-279). IEEE.

[8] Mechee, M. S., Hussain, Z. M., & Salman, Z. I. (2021), "Wavelet Theory," in Wavelet Theory: Applications of the wavelet., Dublin, Ireland, 2021, p. 21.
[9] Iwok, U. U., Udofia, K. M., Obot, A. B., Udofia, K. M., Michael, U. A., & Kingsley, A. I. (2023). Evaluation of Machine Learning Algorithms using Combined Feature Extraction Techniques for Speaker Identification. Journal of Engineering Research and Reports, 25(8), 197-216.
[10] Upadhyaya, P., Farooq, O., & Abidi, M. R. (2018). Mel scaled M-band wavelet filter bank for speech recognition. International Journal of Speech Technology, 21(4), 797-807
[11] Hossain, D., Imtiaz, M. H., Ghosh, T., Bhaskar, V., & Sazonov, E. (2020, July). Real-time food intake monitoring using wearable egocnetric camera. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 4191-4195). IEEE.
[12] Hariri, W. (2022). Efficient masked face recognition method during the covid-19 pandemic. Signal, image and video processing, 16(3), 605-612.
[13] Kumar, K. K., Kasiviswanadham, Y., Indira, D. V., & Bhargavi, C. V. (2023). Criminal face identification system using deep learning algorithm multi-task cascade neural network (MTCNN). Materials Today: Proceedings, 80, 2406-2410.
[14] Atemkeng, M., Osanyindoro V., Rockefeller R., Sisipho H., Jecinta M., Ansah-Narh T., Tchakounté F., & Arnaud N. F. (2024). Ensemble learning and deep learning-based defect detection in power generation plants. Journal of Intelligent Systems, 33(1), 20230283.
[15] Soltani, Z., Sørensen, K. K., Leth, J., & Bendtsen, J. D. (2022). Fault detection and diagnosis in refrigeration systems using machine learning algorithms. International Journal of Refrigeration, 144, 34-45.
[16] Saufi, S. R., Ahmad, Z. A. B., Leong, M. S., & Lim, M. H. (2020). Gearbox fault diagnosis using a deep learning model with limited data sample. IEEE Transactions on Industrial Informatics, 16(10), 6263-6271
[17] Zhu, F., Liu, C., Yang, J., & Wang, S. (2022). An improved MobileNet network with wavelet energy and global average pooling for rotating machinery fault diagnosis. Sensors, 22(12), 4427.
[18] Khaleefah, S. H., Mostafa, S. A., Gunasekaran, S. S., Khattak, U. F., Yaacob, S. S., & Alanda, A. (2024). A deep learning-based fault detection and classification in smart electrical power transmission system. JOIV: International Journal on Informatics Visualization, 8(2), 812-818.
[19] Yoon, D. H., & Yoon, J. (2024). Development of a real-time fault detection method for electric power system via transformer-based deep learning model. International Journal of Electrical Power & Energy Systems, 159, 110069.
[20] Ajimah, E. N., & Iloanusi, O. N. (2024). Biometric voice recognition system in the context of multiple languages: using traditional means of identification of individuals in Nigeria languages and English language. Res. Stat, 2(1), 1-16.
[21] Kumar, S., Goyal, D., Dang, R. K., Dhami, S. S., & Pabla, B. S. (2018). Condition based maintenance of bearings and gears for fault detection–A review. Materials Today: Proceedings, 5(2), 6128-6137.