Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.054886

ARTICLE

Machine Fault Diagnosis Using Audio Sensors Data and Explainable AI Techniques-LIME and SHAP

Aniqua Nusrat Zereen[1], Abir Das[2] and Jia Uddin[3,*]

[1]School of Data and Sciences, Brac University, Dhaka, 1212, Bangladesh

[2]JW KIM College of Future Studies, Endicott College, Woosong University, Daejeon, 300-718, Republic of Korea

[3]AI and Big Data Department, Endicott College, Woosong University, Daejeon, 300-718, Republic of Korea

*Corresponding Author: Jia Uddin. Email: jia.uddin@wsu.ac.kr

Received: 10 June 2024 Accepted: 14 August 2024 Published: 12 September 2024

ABSTRACT

Machine fault diagnostics are essential for industrial operations, and advancements in machine learning have significantly advanced these systems by providing accurate predictions and expedited solutions. Machine learning models, especially those utilizing complex algorithms like deep learning, have demonstrated major potential in extracting important information from large operational datasets. Despite their efficiency, machine learning models face challenges, making Explainable AI (XAI) crucial for improving their understandability and fine-tuning. The importance of feature contribution and selection using XAI in the diagnosis of machine faults is examined in this study. The technique is applied to evaluate different machine-learning algorithms. Extreme Gradient Boosting, Support Vector Machine, Gaussian Naive Bayes, and Random Forest classifiers are used alongside Logistic Regression (LR) as a baseline model because their efficacy and simplicity are evaluated thoroughly with empirical analysis. The XAI is used as a targeted feature selection technique to select among 29 features of the time and frequency domain. The XAI approach is lightweight, trained with only targeted features, and achieved similar results as the traditional approach. The accuracy without XAI on baseline LR is 79.57%, whereas the approach with XAI on LR is 80.28%.

KEYWORDS

Explainable AI; feature selection; machine learning; machine fault diagnosis

1 Introduction

Machine fault diagnostics are essential for industrial operations, and advancements in machine learning (ML) have significantly enhanced these systems by providing accurate predictions and expedited solutions. The models, particularly those employing deep learning (DL) methodologies, have demonstrated the ability to identify potential defects by analyzing data from various controllable variables, such as vibration, temperature, and acoustics. These developments are crucial for predicting faults before they occur, thereby preventing operational disruptions and minimizing costs [1,2].

ML models, especially those utilizing complex algorithms like DL, have demonstrated major potential in extracting important information from large operational datasets [3–5].

Recent studies have underscored the importance of incorporating Explainable AI (XAI) into machine fault diagnosis. Techniques such as shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) are employed to elucidate the decision-making mechanisms utilized by models, enabling the recognition of crucial data features that can predict potential problems, therefore enhancing the understanding of the model's behavior [6–8].

Feature selection is important in enhancing the performance and interpretability of ML models for fault diagnosis. By identifying the most relevant features, feature selection reduces data dimensionality and computational complexity. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Recursive Feature Elimination (RFE) are prominent feature selection techniques used in machine fault diagnosis. The PCA reduces the dimensionality of datasets by transforming original variables into a set of new uncorrelated variables that maximize variance, thus simplifying data interpretation [9]. The LDA improves class separability by finding a linear combination of features that best separates different classes, making it useful for distinguishing fault types [10]. The RFE is an iterative method that removes the least important features based on model performance, often using techniques like support vector machines (SVM) to rank feature importance, ensuring that only the most relevant features are retained [11]. Techniques similar to fault diagnosis in machines using image data play a vital role. Approach to diagnose high-impedance faults, using a combination of semantic segmentation, signal envelope, and Hilbert marginal spectrum, utilizes 1D-UNet for transient process identification in zero-sequence voltage, enhancing the detection of fault inception. The integration of signal envelope and Hilbert marginal spectrum (HMS) and HMS features, transformed into images and analyzed with ResNet18, showcases superior performance in detecting high-impedance faults, particularly in resonant distribution networks. Another research introduces an innovative methodology combining adaptive transient process calibration with multiscale correlation analysis to enhance the accuracy of fault localization [12,13]. Additionally, SHAP and LIME provide insights into model decision-making, enhancing interpretability and trust. The SHAP provides a global perspective by assigning importance values to each feature, revealing how changes in features like vibration or temperature influence the model's predictions. This clarity helps maintenance teams identify critical factors contributing to machine faults and enables preemptive actions. The LIME offers local interpretability by explaining individual predictions, and showing how specific instances of sensor readings lead to certain fault diagnoses. This detailed analysis aids engineers in validating and trusting model decisions on a case-by-case basis. Practical applications, such as pinpointing acoustic signals associated with bearing faults using SHAP or understanding combined temperature and vibration patterns through LIME, demonstrate how these techniques turn complex model outputs into actionable maintenance strategies.

The integration of ML with feature selection techniques in machine fault diagnostics indicates a synergy of technology, strategy, and user-centered design aimed at improving the reliability and efficiency of industrial processes. These technical developments are expected to usher in an era of transformative change in industrial maintenance, marked by major reductions in downtime and continual improvements in operational efficiency [14].

In the proposed methodology for machine fault diagnosis, feature selection is achieved using XAI coupled with Logistic Regression (LR). This process iteratively refines the set of features, enabling the identification and training of the most informative ones, thereby enhancing model accuracy and reducing unnecessary complexity. Initial results indicate that models trained with a targeted selection

of three to five features yield the highest accuracy rates, underscoring the effectiveness of the feature selection strategy.

The approach extends to training various ML algorithms, including Extreme Gradient Boosting (XGBoost), Random Forest (RF), SVM, and LR. This study also explores the potential of other algorithms like Gaussian Naive Bayes (GaussianNB) and Neural Networks for comparative analysis. Crucially, the optimization of hyperparameters is emphasized to maximize each model's performance.

Further refinement is achieved through the application of SHAP to identify the most important features for fault diagnosis. This insight allows for the training of an LR model focused solely on these key features, enhancing diagnostic precision. This methodology is juxtaposed against a traditional approach, which utilizes a predefined set of classifiers evaluated independently. The structured and iterative method highlights the substantial benefits of strategic feature reduction and focused model training in improving fault diagnosis accuracy.

This paper introduces several contributions compared to prior AI-based methods in machine fault diagnosis:

1. *Integration of XAI techniques:* Unlike previous works primarily focusing on black-box models, our study integrates SHAP and LIME to provide transparency in model decision-making. This integration allows operators to understand the rationale behind predictions, enhancing trust and facilitating model acceptance in industrial settings.
2. *Comprehensive feature selection methods*: We employ advanced feature selection techniques such as RFE, PCA, etc. These methods help in identifying the most relevant features, reducing dimensionality, and improving model performance and interpretability. This approach contrasts with earlier studies that may have relied on a limited set of feature selection methods.
3. *Audio sensor data*: The approach leverages audio sensor data for fault diagnosis, a relatively underexplored domain in comparison to traditional vibration or temperature data. This novel use of sound data opens new avenues for fault detection in scenarios where traditional sensors might be less effective.
4. *Enhanced comparative analysis*: We conduct a thorough comparative analysis of state-of-art models and the proposed method, highlighting the performance improvements achieved through the proposed approach. The metrics, such as accuracy, F1 score, precision, and recall are deeply explored, providing a comprehensive evaluation of the model efficacy.

The rest of the paper is organized as follows: Section 2 provides an extensive review of previous works, Section 3 describes the proposed methodology, Section 4 discusses the results, and finally concludes the paper in Section 5.

2 Literature Review

The field of fault diagnosis is critical in maintaining the reliability and efficiency of industrial systems. Accurate fault detection and classification rely on the effective analysis of high-dimensional data, which can be complex and computationally intensive. Data compression techniques, such as PCA, LDA, and Partial Least Squares (PLS) are essential in fault diagnosis systems. The PCA reduces dimensionality, LDA finds linear combinations of features, and PLS identifies fundamental relations between predictors and responses, handling multicollinearity effectively [15]. These techniques improve model performance by maintaining crucial information and enhancing feature space separability. However, PCA assumes linear relationships and may not capture nonlinear interactions, potentially leading to the loss of essential information. The LDA requires normally distributed features

and might underperform with non-linear data. The PLS, while effective with multicollinearity, can become computationally intensive with very large datasets, impacting processing time and resource utilization. Computation increases with the increment of feature dimension. In the proposed approach, we extracted several features from a sample audio file and then we select only the best few features among them which makes the model less computationally burdened.

The XAI elucidates the complexity of diagnostic ML models [16]. The XAI applications can justify AI-driven decisions in industrial environments. Complex models like SHAP and LIME predict behaviors, clarify decision-making, and ensure regulatory compliance. LIME is a prominent example of a surrogate model that provides local approximations to explain individual predictions of a black-box model. It involves generating perturbations of the input data and analyzing the resulting changes in predictions to identify which features are most influential. It helps users understand the decision-making process of complex models on a case-by-case basis [17]. Mean Decrease Accuracy (MDA) is another popular method for feature selection, commonly used in conjunction with Random Forest models. The MDA measures the importance of each feature by assessing the decrease in model accuracy when the feature is permuted. This approach provides an intuitive measure of feature importance, as more critical features will cause a major drop in accuracy when altered [18].

In [19], impulse frequency response analysis-based method employs impulse frequency response analysis combined with image classification using ResNet18 and Smooth Grad-CAM++ to diagnose winding short circuit faults in synchronous machines, achieving high diagnostic accuracy and providing enhanced model interpretability. It utilizes deep learning to analyze complex patterns in image data, making it highly effective for precise fault detection. It shows the importance of feature selection and explainability in ML models for fault diagnosis. However, it stands out for its high accuracy in detecting specific electrical faults using image data, while the proposed approach audio sensor-based, provides a comprehensive framework for fault diagnosis using sound data, supported by robust XAI techniques.

Researchers integrate wavelet weight initialization and adaptive threshold for robust and interpretable fault diagnosis in machines [20,21]. The methods leverage physics-informed models to enhance feature extraction and dynamically adjusts thresholds to improve fault detection accuracy. The studies focus on integrating domain knowledge with ML for better interpretability and robustness, particularly in varying operational conditions. Reference [21] introduces the physics-informed wavelet domain adaptation network designed to improve cross-machine transfer diagnosis by integrating wavelet-based feature extraction with ML. It employs optimized wavelet weights in the first convolutional layer to enhance domain transferability and extract discriminative features. It shows significant performance improvements in challenging cross-machine diagnostic tasks, validating its efficacy across multiple datasets. However, emphasizing real-time audio-based monitoring and explainability is crucial in noisy industrial environments.

In conclusion, the implementation of ML models, multimodal data, and XAI methods in diagnostic and predictive maintenance systems indicates a significant transformation towards more effective and proactive industrial operations. Continuous research and development in these fields will have a major impact on the capabilities of fault diagnostics and predictive maintenance. Table 1 shows the summary of the state-of-the-art models in fault diagnostics.

Table 1: Summary of state-of-the-art fault diagnostics models

Reference	Contribution	Limitation
Al-Kaf et al. [7]	SHAP and LIME for diagnosing open faults in NPC inverters, providing transparency in model decision-making, and enhancing the interpretability of complex ML models in energy conversion systems.	Computationally intensive, imposing remarkable processing power and time; specifically tailored to energy conversion applications, which may limit generalizability to other types of industrial systems.
Begum et al. [8]	User-centric study on the implementation of SHAP and LIME for generating explainable alerts in security operation centers (SOC), highlighting the effectiveness of XAI techniques in improving user trust and understanding in critical security contexts.	Focuses primarily on security operation centers, potentially restricting the applicability of findings to other industrial environments; the approach may not directly address the unique challenges in different sectors.
Islam et al. [9]	Recent advancements and applications in reducing the dimensionality of high-dimensional datasets, which is crucial for simplifying data interpretation and improving model performance in various industrial applications.	Assumes linear relationships among variables, which may result in the loss of crucial information in datasets with nonlinear interactions; may not be suitable for all types of data.
Xanthopoulos et al. [10]	Explanation of LDA for improving class separability in datasets.	Needs normally distributed features, which may limit performance with non-linear data; and may not handle complex, non-linear relationships as effectively as other techniques.
Das et al. [11]	Application of RFE for identifying transformer faults.	The iterative nature of RFE can be computationally intensive and time-consuming, especially with large datasets; it relies heavily on the underlying model's performance, which can be a controlling factor.
Wold [15]	Introduction and application of PLS in handling multicollinearity.	Computationally intensive when applied to very large datasets, which can impact processing time and resource utilization; may need careful tuning and validation to ensure optimal performance.
Breiman [18]	Development of random forests for robust, ensemble-based ML.	Expects large amounts of data; may be computationally expensive.

3 Methodology

The proposed study integrates feature selection with traditional ML techniques, significantly enhancing the performance and interpretability of fault diagnosis models. Key innovations include the use of XAI methods, such as SHAP and LIME, to guide feature selection and provide insights into the decision-making processes of the models. This comprehensive and structured approach to model training focuses on the strategic reduction of features to enhance diagnostic accuracy. The integration of RFE with LR for feature selection stands out as a unique aspect. This method iteratively refines the feature set, enabling the identification of the most informative features, which are then used to train various ML models, including XGBoost, RF, SVM, and GaussianNB. This iterative process, validated through rigorous cross-validation and hyper-parameter tuning, underscores the effectiveness of the feature selection strategy.

We have divided the studies into two parts, which show the difference between the traditional approach and the proposed approach to diagnose faults in machines. We have prepared a specific dataset from the existing large dataset [22]. The traditional approach involves feature extraction of the dataset and training the LR model to classify faults in the machine, whereas the approach does similar feature extraction but chooses the most contributing features among all extracted features and trains the model to classify shown in Fig. 1. This technique shows promising results, which also provide a lightweight model with better accuracy. We also have trained different ML models for comparative analysis, which are XGBoost, SVM, GaussianNB, and RF. The following sections describe the details of the dataset preparation and model training phases. The pseudocode of the proposed model is defined in Algorithm 1.

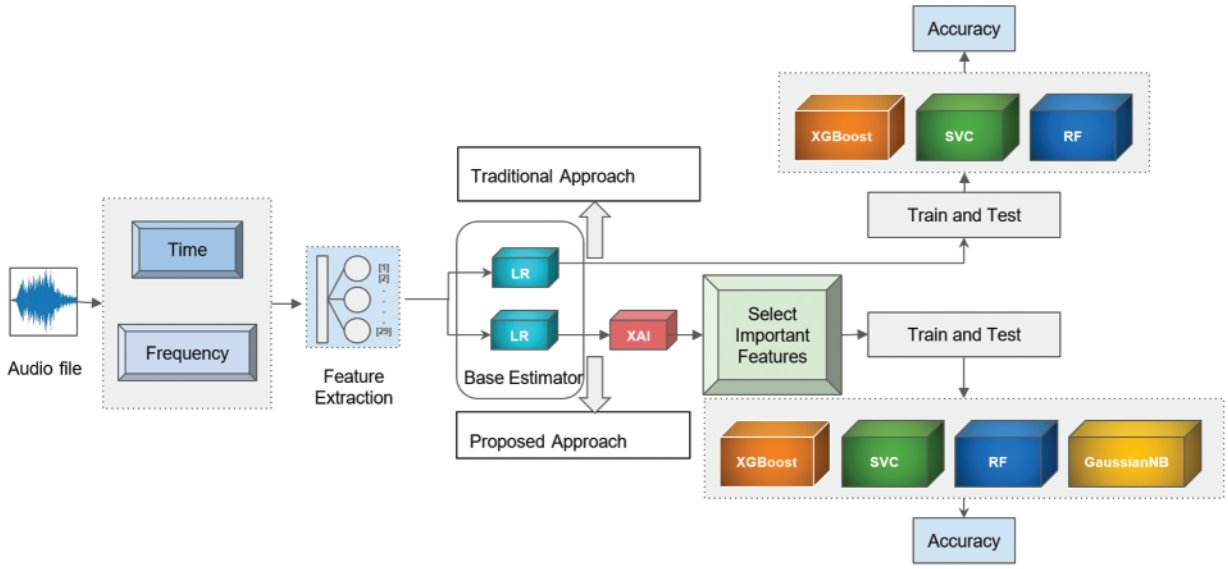

Figure 1: Workflow of the proposed methodology

Algorithm 1:

Input: Audio data from sensors, labels (normal/abnormal)
Output: Trained model, performance metrics, and explanations
1: Load and preprocess audio data
2: Initialize feature extraction functions
3: Initialize classifiers: LR, XGBoost, SVC, GaussianNB, RF
4: Initialize explainable AI tools: LIME, SHAP
Feature Extraction:
5: for each audio sample x in dataset do
6: time_features ← extract_time_domain_features(x)
7: freq_features ← extract_frequency_domain_features(x)
8: features ← time_features + freq_features
9: Append features to feature matrix X
10: Append label to label vector y
11: end for
Model Training and Evaluation:
12: Split data into training set and test set
13: Scale features using StandardScaler
14: for each classifier in classifiers do
15: Train classifier on training set
16: Predict on test set
17: Calculate performance metrics: accuracy, F1 score, precision, recall
18: Plot confusion matrix
19: end for
ROC Curve:
20: Plot ROC curves for all classifiers
Explainable AI:
21: Select an instance from test set for explanation
22: LIME explanation ← explain_instance(LIME, instance, classifier)
23: Display LIME explanation
24: SHAP explanation ← SHAP values for X_test
25: Plot SHAP summary
26: Return trained models, performance metrics, LIME, and SHAP explanations

3.1 Data Collection and Preprocessing

Data collection and preprocessing are critical steps in the methodology of this study, impacting the quality and reliability of the machine fault diagnosis models. The audio recordings used in this research were sourced from industrial fans, with the dataset comprising two distinct sound categories: 'normal' and 'abnormal' [22]. This dataset was extracted from a larger repository of machine sounds, ensuring a comprehensive representation of operational conditions.

Each audio file in the dataset is 30 s long and saved in the waveform audio file (wav) format. The total dataset includes 1514 audio files, with 1107 classified as normal and 407 as abnormal. This distribution allows for a robust analysis of fault detection and diagnosis.

Preprocessing involves transforming raw audio signals into a set of 29 informative features, derived from both time and frequency domains. Time domain features include statistical measures like mean, median, variance, and standard deviation, which provide insights into the central tendency and dispersion of the audio signals. Additional metrics, such as skewness, kurtosis, and zero-cross rate offer further characterization of the signal distribution and oscillatory behavior.

In the frequency domain, features, such as spectral centroid, bandwidth, and spectral contrast are extracted, providing a detailed frequency analysis of the audio signals. This dual-domain approach ensures a comprehensive feature set that captures both temporal and spectral characteristics of the sound data.

3.2 Feature Extraction

Feature extraction is crucial in diagnosing machine faults using sound or vibration data. The approach transforms the raw audio signal into 29 informative features, both from time and frequency domains, as listed in Table 2.

Table 2: Time domain and frequency domain features

Time domain features	Frequency domain features
Mean	Mean_freq
Median	Median_freq
Variance	Variance_freq
Std_dev	Std_dev_freq
Skewness	Skewness_freq
Kurt	Kurt_freq
Zero_cross_rate_value	Delta
Num_waves	Alpha
Wave_duration	Beta
Inst_freq	Gamma
Mobility	Sigma
Activity	Theta
Complexity	Zero_a
Energy	

3.2.1 Time Domain Features

The time domain features include statistical measures, such as the mean, median, variance, and standard deviation (Std_dev), which provide insights into the central tendency and dispersion of the audio signal. Additionally, skewness and kurtosis (Kurt) metrics capture the asymmetry and tailedness of the signal distribution, respectively. The Zero_cross_rate_value indicates the rate of sign changes in the audio signal, while num_waves and wave_duration quantify the oscillatory behavior and duration of waveforms. Instantaneous frequency (Inst_freq) and mobility reflect the rate of phase change and signal variation, whereas activity and energy denote the signal's total energy and variance-related energy. The complexity measure assesses the dynamic changes in the signal's frequency content, and the k-complex identifies specific waveform patterns.

3.2.2 Frequency Domain Features

In the frequency domain, features such as mean frequency (Mean_freq), median frequency (Median_freq), variance frequency (Variance_freq), and Standard Deviation of Frequency (Std_dev_freq) offer a detailed understanding of the signal's frequency components. The Skewness_freq and Kurt_freq further characterize the distribution of these frequencies. Power or amplitude within specific frequency bands is quantified through features like delta, alpha, beta, gamma, sigma, and theta, each relevant to different operational states and potential fault conditions of the fan. The Zero_a feature represents the rate of zero crossings in the frequency domain, and the b_a ratio (beta to alpha power) provides additional insights into the operational health of the fan.

The following equations provide a detailed understanding of the features, helping in the accurate and interpretable diagnosis of machine faults [23]:

$$Mean = \frac{1}{N}\sum_{i=1}^{N} x[i] \tag{1}$$

$$Median = x\left[\frac{N+1}{2}\right] \tag{2}$$

$$\text{Variance} = \frac{1}{N}\sum\nolimits_{i=1}^{N} (x[i] - Mean)^2 \tag{3}$$

$$Std_{dev} = \sqrt{Variance} \tag{4}$$

$$Skewness = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x[i] - Mean}{Std\_dev}\right)^3 \tag{5}$$

$$\text{Kurtosis} = \frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\frac{x[i] - Mean}{Std\_dev}\right)^4 - 3 \tag{6}$$

$$Zero\_cross\_rate\_value = \frac{1}{N}\sum_{i=1}^{N-1} 1_{\{x[i].x[i+1]<0\}} \tag{7}$$

$$Num\_waves = Number\ of\ Zero - crossing \tag{8}$$

$$Wave\_duration = \frac{Total\ duration\ of\ signal}{Num\_waves} \tag{9}$$

$$Inst\_freq = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{10}$$

$$Mobility = \frac{\text{Std\_dev of the first derivative of the signal}}{\text{Std\_dev of the signal}} \tag{11}$$

$$Activity = Variance \tag{12}$$

$$Complexity = \frac{Mobility\ of\ the\ first\ derivative\ of\ the\ signal}{Mobility\ of\ the\ signal} \tag{13}$$

$$Energy = \sum_{i=1}^{N} x[i]^2 \tag{14}$$

$$Zero\_a = \frac{1}{N-1} \sum_{i=1}^{N-1} 1_{\{f[i].f[i+1]<0\}} \tag{15}$$

$$b\_a = \frac{Power\ in\ beta\ band}{Power\ in\ alpha\ band} \tag{16}$$

3.3 Feature Scaling

Before training the ML models, feature scaling was applied as a crucial pre-processing step to normalize the feature values and enhance the convergence properties of the learning algorithms. Standard scaling, also known as Z-score normalization, was employed for this purpose. We compute the mean and standard deviation of each feature of the training data and scale the features accordingly. We used the same steps for test data to ensure that the scaling parameters were consistent between the training and testing datasets.

We standardize the range of feature values using standard scaling which prevents features with larger magnitudes from dominating the learning algorithm during model training. Standard scaling involves transforming the feature values such that they have a mean of *0* and a standard deviation of *1*. This is achieved by subtracting the mean of each feature from its value and then dividing by the standard deviation of the feature. This ensures that all features contribute equally to the learning process, leading to more stable and efficient training of ML models. The formula for standard scaling is shown in Eq. (17) as:

$$Xscaled = \frac{x - \mu}{\sigma} \tag{17}$$

where x is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature.

3.4 Model Selection

The methodology utilizes a multi-pronged approach, exploring the capabilities of different algorithms to identify the most effective model for the specific fault diagnosis task. A foundational aspect of this study is the selection of an appropriate baseline classifier, which serves as a reference for comparative analysis. We utilized LR for its inherent simplicity and effectiveness in binary classification problems. It provides a straightforward linear model that effectively captures the relationships between the extracted features and the target classes, making it a suitable choice for initial model evaluation and feature relevance analysis.

Beyond LR, a diverse array of classifiers, including XGBoost, SVM, GaussianNB, and RF, are evaluated. Each of these models offers unique advantages and complexities, ranging from the ensemble-based learning of RF to the kernel-based decision boundaries of SVM. XGBoost, known for its gradient-boosting framework, provides robust performance by combining the predictions of several weak models to produce a powerful ensemble. The SVM, with its capability to handle high-dimensional spaces and its versatility with different kernel functions, excels in scenarios where clear margin separation is crucial. GaussianNB, with its probabilistic approach, offers simplicity

and computational efficiency, particularly effective when the assumption of feature independence holds true.

Through rigorous experimentation and cross-validation, these classifiers are trained using the 29 features derived from the training dataset, aiming to recognize their respective capabilities in classifying abnormal and normal machine operations. This traditional approach serves as a benchmark for evaluating the performance improvements achieved through feature selection techniques.

In the proposed approach, the entire process mentioned above is repeated but with an emphasis on using the most important features in the training phase. The important features are selected using RFE, a method that recursively removes the least significant features based on the model's performance until the optimal feature subset is identified. This targeted feature selection enhances the model's efficiency and accuracy by focusing on the most relevant data attributes.

To accurately assess the performance of the trained ML models, the dataset was carefully divided into separate training and testing subsets. For model training, we utilize 80% of the data, while the remaining 20% is used for model evaluation. To ensure reproducibility and robustness in the model evaluation process, a random seed is assigned as a deterministic factor that controls the pseudo-random division of the dataset. This ensures that the training and testing sets are consistently generated, allowing for reliable performance comparisons across different models and feature selection strategies.

3.4.1 Hyper-Parameter Tuning

Each chosen ML algorithm has its own set of hyper-parameters that control its learning behavior. We tuned the hyper-parameters of these models empirically to achieve better performance. In addition, we employ grid search or randomized search techniques to explore a predefined range of values for each hyper-parameter. To mitigate overfitting and ensure the model generalizes well to unseen data, we employ a cross-validation strategy.

3.4.2 Feature Selection with Recursive Feature Elimination

While we extract a comprehensive set of features, not all features may be equally important for accurate fault diagnosis. The RFE helps identify the most relevant features. We utilize LR as the base estimator for RFE. First, we train LR on the entire 29-feature set. Then the RFE iteratively removes the feature with the least contribution to the model's performance, as determined by the base estimator's feature importance scores. After that, we retrain the LR on the reduced feature set. Focusing on the most informative features can potentially lead to better classification accuracy and reduce the risk of over-fitting.

In addition, a smaller set of relevant features makes it easier to understand the model's decision-making process.

3.5 Evaluation Metrics

We use a set of standard performance metrics to evaluate the classification effectiveness of the trained models. These metrics provide quantitative insights into the model's ability to accurately distinguish between normal and abnormal machine operations.

Accuracy measures the overall proportion of correct predictions made by the model, including both true positives and true negatives. The formula for accuracy is [24]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{18}$$

where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

Precision assesses the accuracy of the positive predictions made by the model. The formula for precision is:

$$\text{Precision} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP}} \tag{19}$$

Recall measures the model's ability to detect all actual positives. The formula for the recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

The F1-score is the mean of precision and recall, providing a balance between the two. It is particularly useful when the class distribution is uneven. The formula for the F1-score is:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

4 Experimental Results Analysis

In this section, we present the experimental results obtained from the machine fault diagnosis study using two distinct approaches: the proposed approach and the traditional approach to classifying instances of abnormal and normal machine operations based on extracted features.

4.1 Traditional Approach without XAI

The traditional approach employs a set of predefined ML classifiers, including SVC, XGBoost, LR, and RF. Each classifier is trained on 29 features and evaluated independently, as shown in Table 3. The SVC model achieved the highest recall at 100%, indicating its superior ability to identify all relevant cases. However, the RF model outperformed the others in terms of F1-score, Precision, and Accuracy, with scores of 89.65%, 81.88%, and 82.04%, respectively, suggesting its overall effectiveness and balance in prediction accuracy and reliability among the evaluated models. The XGBoost and LR models showed competitive but slightly lower performance metrics in comparison.

Table 3: Model performance with the traditional approaches

Model	Recall	F1-score	Precision	Accuracy
XGBoost	97.32	88.54	80.00	80.63
SVC	100.00	88.12	78.02	79.93
LR	99.02	87.32	78.87	79.57
RF	99.54	89.65	81.88	82.04

4.2 Proposed Approach with XAI

As a base estimator, we use LR to determine the most important features contributing to the model. We utilize SHAP to show the distribution of SHAP values for each feature in the model. The features are listed on the *y*-axis, and the SHAP value distribution is shown on the *x*-axis. The color of the distribution indicates the impact of the feature on the model's output. Blue indicates a negative impact, and red indicates a positive impact. The force of the color indicates the magnitude of the impact shown in Fig. 2.

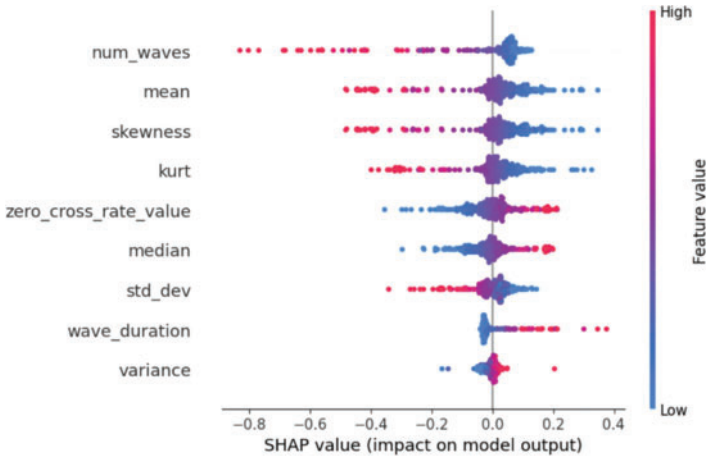

Figure 2: Visualization of the impact of the top nine features on a logistic regression model using SHAP

The features 'mean', 'skewness', 'wave_duration', and 'std_dev' all influence the model's predictions in different ways. The 'mean' has a positive effect, shown by its reddish color, but it's not as strong as the effect from 'num_waves'. The 'skewness' also has a positive effect but it's even weaker than the 'mean'. On the other hand, 'wave_duration' tends to lower the model's predictions, which is indicated by its blue color. The 'std_dev' also lowers predictions but not as much as 'wave_duration'. Some features do not impact the model's predictions much at all. These features show colors close to zero on the *x*-axis, meaning they do not change the predictions much either way. The spread of colors for each feature tells us how consistently affected the model. The 'num_waves' has a narrow red spread, meaning it usually similarly increases predictions across different data points. The 'wave_duration' has a narrow blue spread, showing it consistently lowers predictions. However, 'mean' and 'skewness' have wider color spreads, meaning their effects on the predictions can vary a lot depending on the specific data point.

The proposed approach begins with feature selection using XAI in conjunction with LR. We iteratively select a varying number of the top nine contributing features and train LR models to evaluate their performance. The results of this experiment, summarized in Table 4, demonstrate the accuracy of LR models trained with different numbers of selected features.

Table 4: Accuracy based on number of selected features

No. of features	3	4	5	6	7	8	9
Model accuracy	80.28	79.93	80.28	80.28	79.93	79.93	79.93

Fig. 3 shows that adding more features usually improves an ML model's accuracy, but only up to a certain point. After this point, adding more features does not help and might even make the model less accurate. For this specific model, the graph indicates that the best performance happens when using six features. After adding more than six features, the accuracy drops. In addition, we utilize LIME to explain the prediction for a single instance of the model and consider all features that contribute to that prediction. Fig. 4 shows how the features together contribute to the model's prediction for a particular instance. The instance is the result of the classification between abnormal and normal audio files. The model predicts with a 0.95 probability that the sound is abnormal. The feature influencing this classification the most is 'skewness_freq', with a value of 3.75. Similarly, the model predicts with a 0.05 probability that the sound is normal. The feature influencing this classification the most is 'kurt_freq', with a value of 4.17.

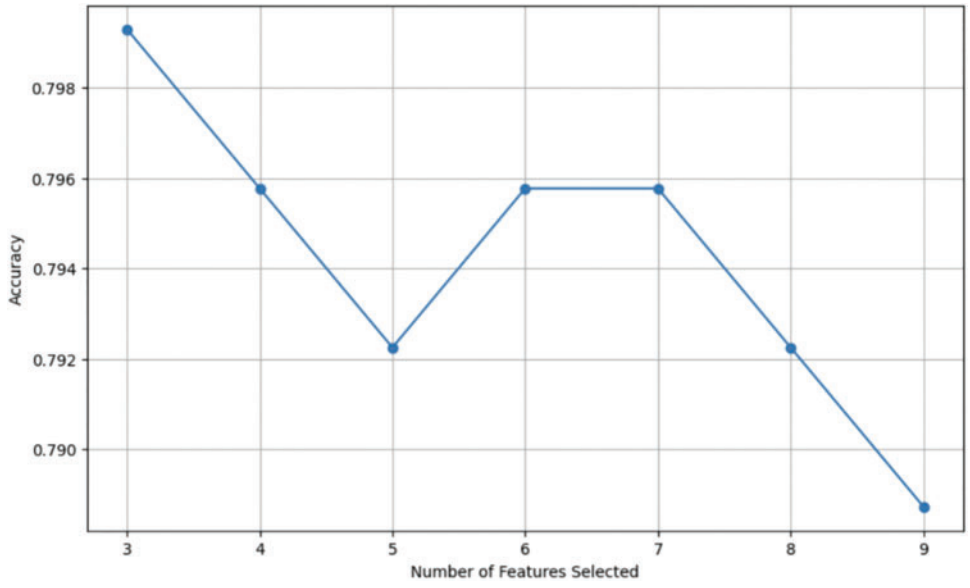

Figure 3: Accuracy curve of the model on a selected number of features

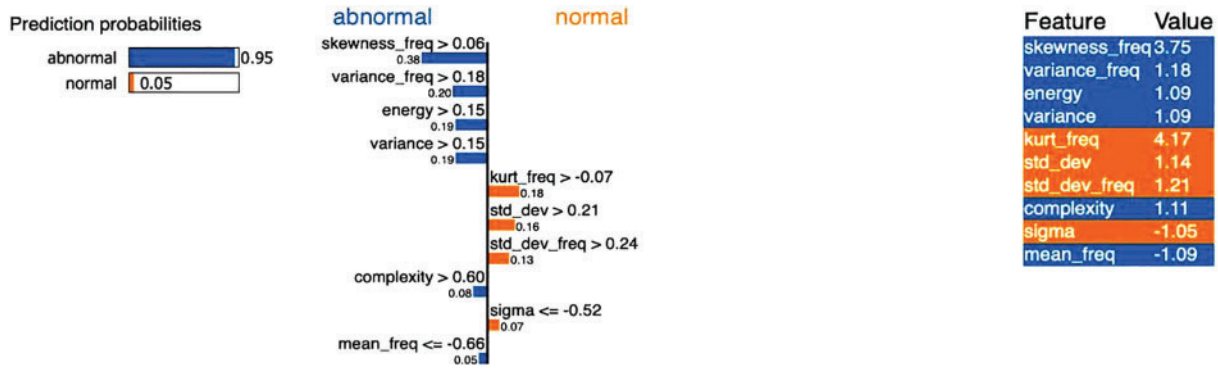

Figure 4: Visualization of the impact of the features on a single instance of the LR model using LIME

Following feature selection, we evaluate the performance of various ML algorithms, including XGBoost, SVC, GaussianNB, and RF. Table 5 shows the accuracy, F1-score, and confusion matrix for each model.

Table 5: Model performance on selected features

Model	Recall	F1-score	Precision	Accuracy
XGBoost	81.33	79.24	81.72	81.33
SVC	71.47	59.94	79.64	71.47
GaussianNB	79.22	76.13	79.91	79.22
RF	82.04	79.37	84.35	82.04

The XGBoost model exhibited the highest accuracy of 81.33% among all the algorithms tested. However, it is worth noting that the SVC model showed relatively lower accuracy compared to the other algorithms, achieving 71.48%. Overall, the proposed approach demonstrated competitive performance, particularly with the LR model and XGBoost classifier.

In the Receiver Operating Characteristic (ROC) curve, the x-axis represents the False Positive Rate (FPR), which is the proportion of negative instances incorrectly classified as positive. The y-axis represents the True Positive Rate (TPR), which is the proportion of positive instances correctly classified. A perfect classifier would classify all positive instances correctly (TPR $= 1$) and have no false positives (FPR $= 0$). This is represented by the top left corner of the graph. The diagonal line (dashed line in the image) represents a classifier with no discriminative power–it essentially guesses randomly. The area under the ROC curve (Area under the ROC Curve (AUC)) is a numerical measure of a classifier's performance. A larger AUC indicates better performance. In Fig. 5, the RF classifier has the largest AUC (0.79) which means it has the best overall performance among the classifiers displayed.

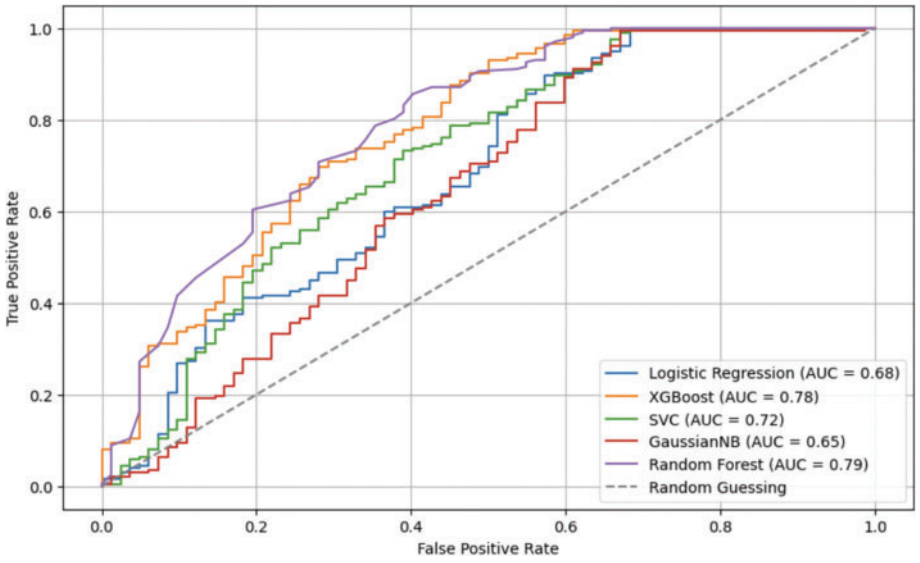

Figure 5: ROC curves of different classifiers

Table 6 shows a summary of the performance metrics: precision, recall, and F1-score for different machine learning models trained with three feature selection techniques: PCA, LDA, and RFE.

Table 6: Performance metrics of ML models trained with three feature selection techniques

Model	Feature selection	Precision (Class 0)	Recall (Class 0)	F1-score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-score (Class 1)	Accuracy
LR	PCA	0.93	0.32	0.47	0.78	0.99	0.87	0.80
LR	LDA	0.90	0.32	0.47	0.78	0.99	0.87	0.79
LR	RFE	0.87	0.32	0.46	0.78	0.98	0.87	0.79
XGBoost	PCA	0.66	0.38	0.48	0.78	0.92	0.85	0.76
XGBoost	LDA	0.76	0.39	0.52	0.79	0.95	0.86	0.79
XGBoost	RFE	0.64	0.44	0.52	0.80	0.90	0.85	0.77
SVC	PCA	1.00	0.30	0.47	0.78	1.00	0.88	0.80
SVC	LDA	1.00	0.32	0.47	0.78	1.00	0.88	0.80
SVC	RFE	0.93	0.33	0.49	0.78	0.99	0.85	0.80
GaussianNB	PCA	0.68	0.34	0.46	0.78	0.94	0.85	0.76
GaussianNB	LDA	0.76	0.32	0.45	0.78	0.96	0.86	0.77
GaussianNB	RFE	0.76	0.35	0.48	0.78	0.96	0.86	0.78
RF	PCA	0.80	0.33	0.48	0.78	0.97	0.86	0.79
RF	LDA	0.48	0.48	0.48	0.79	0.79	0.79	0.70
RF	RFE	0.79	0.38	0.51	0.79	096	0.87	0.79

- LR: Across all feature selection methods, LR consistently shows high precision and recall for Class 1 (positive class), indicating robust classification performance.
- XGBoost: Generally, performs well with PCA and LDA, achieving balanced precision and recall metrics for both classes.
- SVC: Achieves perfect precision for Class 0 with PCA and LDA, suggesting potential overfitting or high sensitivity to feature selection.
- GaussianNB: Shows balanced performance metrics across different feature selection methods, indicating robustness to varying feature subsets.
- RF: Demonstrates competitive performance, especially with RFE, which consistently improves recall for Class 0 while maintaining high metrics for Class 1.

While the current results highlight the impact of different feature selection techniques on model performance, integrating SHAP and LIME techniques could enhance model interpretability further. The SHAP and LIME provide insights into feature importance and local explanations, respectively, aiding in understanding model decisions and improving trustworthiness in practical applications.

In conclusion, leveraging advanced feature selection methods alongside SHAP and LIME techniques could lead to more interpretable and reliable ML models, facilitating informed decision-making in various domains.

4.3 Comparative Analysis

To provide a comprehensive comparison between the proposed and the traditional approaches, we evaluated the performance of the LR and XGBoost classifiers, which were common to both approaches. The LR model achieved an accuracy of 80.28% in the proposed approach, whereas the XGBoost classifier attained an accuracy of 80.63% in the traditional approach. Although the XGBoost classifier outperformed the LR model marginally, both approaches yielded comparable results, demonstrating the effectiveness of the LR model in feature selection. The proposed approach achieved an accuracy ranging from 71.47% to 82.04%, whereas the traditional approach achieved an accuracy ranging from 79.57% to 82.04%. The proposed approach demonstrates that selecting important features using XAI can achieve accuracy close to the traditional approach of selecting all features, which yields a lightweight model approach.

Building on the comparative analysis between the traditional approach and the proposed method, this section delves deeper into the performance metrics of the evaluated models. By analyzing the confusion matrices for each model (as shown in Fig. 6), we can shed light on the reasons behind variations in accuracy, F1-score, precision, and recall.

- LR: While achieving a decent accuracy (0.796), the model struggles with false positives (56). This suggests the model might be overly sensitive and classify negative instances as positive.
- XGBoost: Similar to LR, XGBoost exhibits a high number of FP (49) despite acceptable accuracy (0.806). This indicates potential overfitting to the training data.
- SVM: Like the previous models, SVM has a high FP rate (57) with moderate accuracy (0.799). This suggests further optimization might be required to improve its ability to distinguish between positive and negative classes.
- GNB: The GNB model shows a balance between TP (29) and FP (53). However, its lower accuracy (0.771) compared to other models suggests room for improvement in overall classification performance.
- RF: The RF model demonstrates the best performance among the evaluated models with a high number of TP (32) and low FP (50). This translates to good accuracy (0.809) and balanced precision and recall.

The analysis of the metric variations is as follows:

- Accuracy: While all models achieved moderate accuracy, RF emerged as the most accurate classifier. This suggests the effectiveness of the method (potentially used in RF) in achieving a better balance between TP and negatives.
- F1-score: The F1-score variations reflect the trade-off between precision and recall. RF again exhibits a superior F1-score (0.782), indicating a good balance between identifying true positives and avoiding FP.
- Precision: The high FP rates in LR, XGBoost, and SVM lead to lower precision compared to RF. This means these models classified many negative instances as positive, impacting the precision of their positive classifications.
- Recall: All models achieved acceptable recall, with RF having the highest (0.809). This indicates they were successful in identifying a good portion of the actual positive cases. However, models with high false positives might achieve high recall due to overclassifying negative instances as positive.

By analyzing the confusion matrices and performance metrics, we can see that the method, potentially implemented in the RF, offers a significant improvement in terms of reducing FP while

maintaining good overall accuracy and balanced precision and recall. Further investigation into the specific techniques used in a method can be conducted to pinpoint the factors contributing to this superior performance.

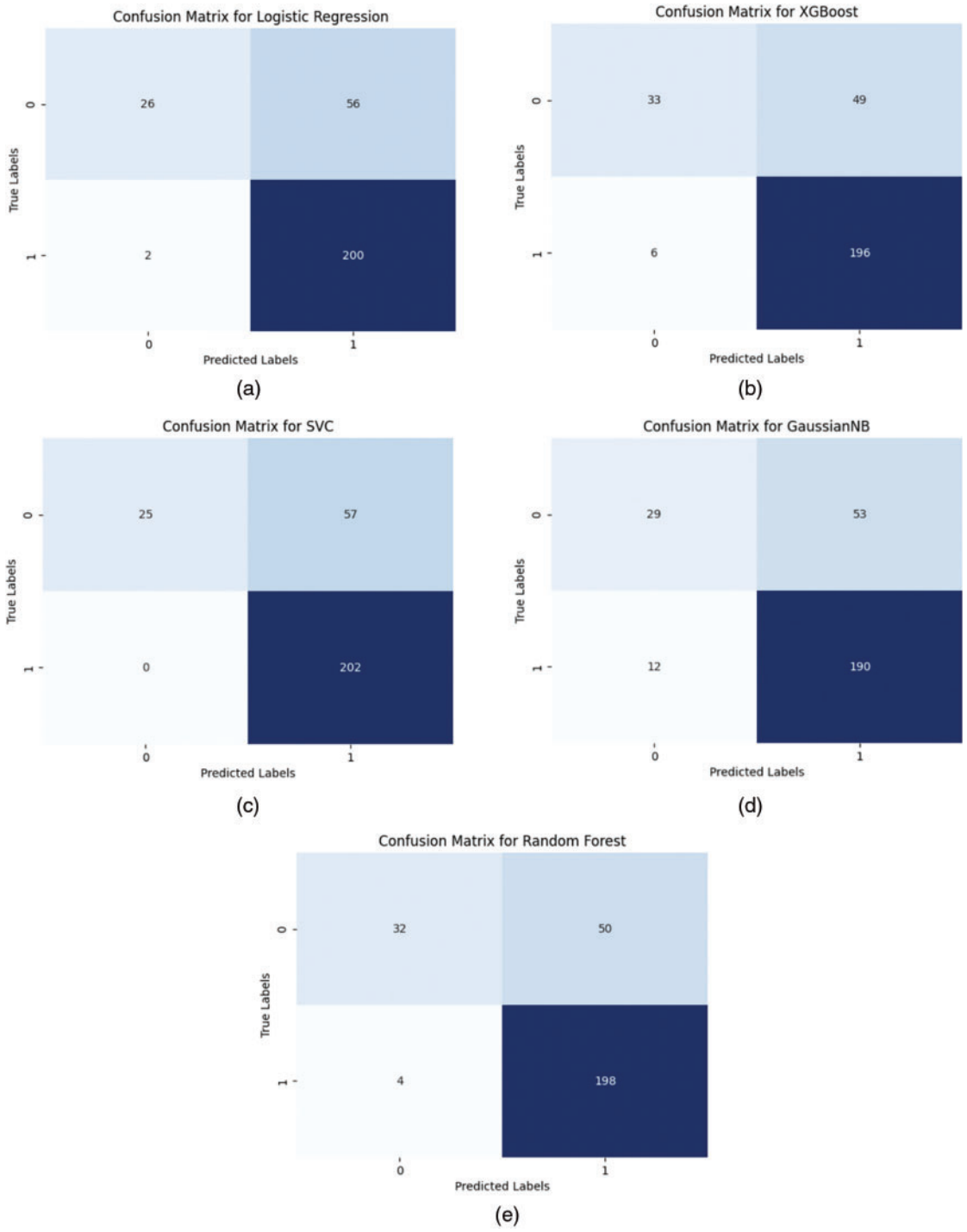

Figure 6: Confusion metric of (a) LR, (b) XGBoost, (c) SVC, (d) GaussianNB, and (e) RF

4.4 t-SNE Visualization of Test Data

To gain deeper insights into the distribution and separability of the test data, we employed t-distributed Stochastic Neighbor Embedding (t-SNE), a powerful technique for visualizing high-dimensional data in a lower-dimensional space. The resultant t-SNE plot, as illustrated in Fig. 7, demonstrates several noteworthy observations regarding the structure and class distribution of the dataset.

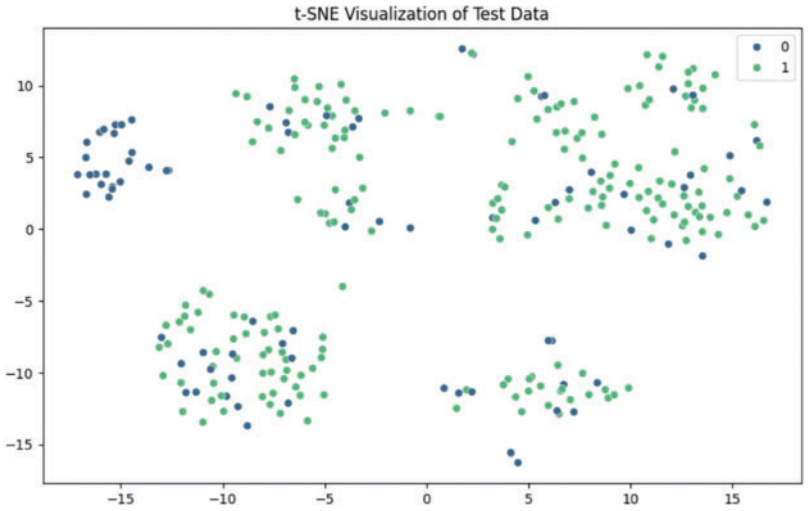

Figure 7: 0 (blue) represents abnormal and 1 (green) represents normal

The t-SNE plot reveals the presence of distinct clusters within the data, indicating that t-SNE has effectively preserved the local and global structures during the dimensionality reduction process. These clusters suggest that the underlying features used in our model capture meaningful patterns and inherent groupings within the data. Notably, while some clusters exhibit clear boundaries, others display varying degrees of overlap, which may pose challenges for classification algorithms.

Data points in the t-SNE visualization are color-coded based on their respective classes, with class 0 represented in blue (abnormal) and class 1 in green (normal). This color-coding highlights several key aspects:

- Certain regions of the plot are dominated by a single class, suggesting that in these regions, the classifier can easily differentiate between the classes.
- Conversely, there are areas with significant overlap between the two classes, indicating regions where the classifier may struggle to achieve high accuracy due to the similarity in feature space.

The t-SNE visualization serves as an intuitive tool for understanding the high-dimensional test data's structure and class separability. By reducing the dimensionality to two dimensions, t-SNE facilitates an accessible interpretation of the complex relationships between data points, aiding in diagnosing potential issues in both the data and the classification models. This visualization underscores the importance of feature selection and the potential need for more sophisticated models to address regions of class overlap.

The t-SNE plot provides valuable insights into the clustering tendencies and class distribution within the test dataset. This analysis is instrumental in understanding the strengths and limitations of the classification model, guiding further refinement of the feature selection and modeling strategies.

5 Conclusion

The proposed study demonstrates the significant potential of integrating feature selection techniques with traditional ML methods, enhanced by XAI techniques such as SHAP and LIME, for machine fault diagnosis using audio sensor data. Our approach not only enhances diagnostic accuracy but also provides valuable insights into the decision-making processes of the models, thereby improving interpretability and trustworthiness. The utilization of audio sensor data for fault diagnosis presents a novel and complementary approach to traditional methods based on vibration or temperature data. However, the approach assumes high-quality input data with effective preprocessing and sufficient representative datasets for accurate model training. Limitations include potential challenges in environments with heavily contaminated audio data and the reliance on high-quality labeled data. The methodology is particularly suitable for industrial environments requiring continuous machinery monitoring and scenarios where understanding the model's decision-making process is critical. Moreover, the comparative analysis highlighted the trade-offs between different ML algorithms and the importance of selecting appropriate algorithms based on the specific features of the dataset.

In conclusion, the experimental results suggest that a systematic approach combining feature selection with ML algorithms can improve the accuracy and efficiency of machine fault diagnosis systems. Both the proposed and traditional approaches demonstrated effectiveness in machine fault diagnosis. Overall, while the traditional approach may provide slightly higher accuracy, the proposed approach provides a simpler and more interpretable solution, which could be advantageous in certain scenarios.

Acknowledgement: The authors appreciate it that this research is funded by Woosong University Academic Research 2024.

Funding Statement: This research is funded by Woosong University Academic Research 2024.

Author Contributions: Aniqua Nusrat Zereen: writing draft, visualization; Abir Das: coding, result analysis, visualization; Jia Uddin: supervisor, conceptual, review, fund acquisition. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: To validate the model, we used the following public dataset: https://zenodo.org/records/4740355, https://zenodo.org/records/4740355 (accessed on 10 January 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Chen, C. Li, and R. -V. Sanchez, "Gearbox fault diagnosis based on deep recurrent neural networks," *J. Manuf. Syst.*, vol. 54, pp. 93–103, 2020.

[2] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 18, no. 10, 2018, Art. no. 3481.

[3] V. Rajan, N. Sobhana, and R. Jayakrishnan, "Machine fault diagnostics and condition monitoring using augmented reality and IoT," *Mech. Syst. Signal. Process.*, vol. 112, pp. 273–284, 2018. doi: 10.1109/ICCONS.2018.8663135.
[4] O. Das and D. B. Das, "Smart machine fault diagnostics based on fault specified discrete wavelet transform," *J. Braz. Soc. Mech. Sci. Eng.*, vol. 45, pp. 519–532, 2023. doi: 10.1007/s40430-022-03975-0.
[5] R. Sewada, A. Jangid, P. Kumar, and N. Mishra, "Explainable Artificial Intelligence (XAI)," *Int. J. Food Nutr. Sci.*, vol. 10, pp. 1–15, 2023. doi: 10.36893/JNAO.2022.V13I02.041-047.
[6] A. M. A. Salih *et al.*, "Commentary on explainable artificial intelligence methods: SHAP and LIME," *Adv. Intell. Syst.*, pp. 1–8, 2024. doi: 10.1002/aisy.202400304.
[7] H. A. G. Al-Kaf and K. -B. Lee, "Explainable machine learning method for open fault detection of NPC inverter using SHAP and LIME," in *IEEE Conf. Energy Convers. (CENCON)*, Kuching, Malaysia, 2023, pp. 14–19.
[8] M. Begum, M. H. Shuvo, I. Ashraf, A. Al Mamun, J. Uddin and M. A. Samad, "Software defects identification: Results using Machine Learning and Explainable Artificial Intelligence Techniques," *IEEE Access*, vol. 11, pp. 132750–132765, 2023. doi: 10.1109/ACCESS.2023.3329051.
[9] M. R. Islam, J. Uddin, and J. M. Kim, "Acoustic emission sensor network based fault diagnosis of induction motors using a gabor filter and multiclass support vector machines," *Adhoc Sens. Wireless Netw.*, vol. 34, pp. 273–287, 2016.
[10] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos and T. B. Trafalis, "Linear discriminant analysis," *Robust Data Min.*, pp. 27–33, 2013. doi: 10.1007/978-1-4419-9878-1.
[11] S. Das, A. Paramane, S. Chatterjee, and U. Rao, "Accurate identification of transformer faults from dissolved gas data using recursive feature elimination method," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 30, pp. 466–473, 2023. doi: 10.1109/TDEI.2022.3215936.
[12] J. H. Gao, M. F. Guo, S. Lin, and D. Y. Chen, "Application of semantic segmentation in high-impedance fault diagnosis combined with signal envelope and hilbert marginal spectrum for resonant distribution networks," *Expert. Syst. Appl.*, vol. 231, 2023, Art. no. 120631. doi: 10.1016/j.eswa.2023.120631.
[13] J. H. Gao, M. F. Guo, S. Lin, and D. Y. Chen, "Advancing high impedance fault localization via adaptive transient process calibration and multiscale correlation analysis in active distribution networks," *Measurement*, vol. 229, 2024, Art. no. 114431. doi: 10.1016/j.measurement.2024.114431.
[14] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Inf. Fusion*, vol. 99, 2023, Art. no. 101805. doi: 10.1016/j.inffus.2023.101805.
[15] H. Wold, "Partial least squares," *Encycl. Stat. Sci.*, pp. 1–19, 1985. Accessed: Mar. 1, 2024. [Online]. Available: https://pure.iiasa.ac.at/id/eprint/2336/1/CP-83-046.pdf
[16] M. F. Siddique, Z. Ahmad, N. Ullah, and J. -M. Kim, "A hybrid deep learning approach: Integrating short-time Fourier Transform and continuous wavelet transform for improved pipeline leak detection," *Sensors*, vol. 23, no. 19, 2023, Art. no. 8079. doi: 10.3390/s23198079.
[17] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. 2022 ACM Conf. Fairness, Account., Transparency*, New York, NY, USA, 2022, pp. 2239–2250.
[18] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.
[19] Y. Chen, Z. Zhao, Y. Yu, W. Wang, and C. Tang, "Understanding IFRA for detecting synchronous machine winding short circuit faults based on image classification and smooth Grad-CAM++," *IEEE Sens. J.*, vol. 23, no. 3, pp. 2422–2432, 2022. doi: 10.1109/JSEN.2022.3225210.
[20] H. Chao, H. Shi, X. Liu, and J. Li, "Physics-informed interpretable wavelet weight initialization and balanced dynamic adaptive threshold for intelligent fault diagnosis of rolling bearings," *J. Manuf. Syst.*, vol. 70, pp. 579–592, 2023. doi: 10.1016/j.jmsy.2023.08.014.
[21] H. Chao, H. Shi, X. Liu, and J. Li, "Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis," *Knowl.-Based Syst.*, vol. 288, 2024, Art. no. 111499. doi: 10.1016/j.knosys.2024.111499.

[22] R. Tanabe *et al.*, "MIMII Due: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," in *Proc. 2021 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, 2021, pp. 21–25.

[23] X. Wang, Y. Zheng, Z. Zhao, and J. Wang, "Bearing fault diagnosis based on statistical locally linear embedding," *Sensors*, vol. 15, no. 7, pp. 16225–16247, 2015. doi: 10.3390/s150716225.

[24] M. Zabin, H. J. Choi, and J. Uddin, "Hybrid deep transfer learning architecture for industrial fault diagnosis using Hilbert transform and DCNN-LSTM," *J. Supercomput.*, vol. 79, no. 5, pp. 5181–5200, 2023. doi: 10.1007/s11227-022-04830-8.