Machine Fault Diagnosis Using Random Forest with Recursive Feature Elimination and Cross Validation

[1]Vetrithangam D, [2]Shamik Palit, [3]Anshu Mehta, [4]Gaddam Saranya, [5]Donamol Joseph and [6]Abhinav Pathak
[1,3]Department of Computer Science & Engineering, Chandigarh University, Punjab, India.
[2]Department of Computing Science and Software Engineering, University of Stirling RAK Campus, Al Dhait South, Ras Al Khaimah, United Arab Emirates.
[4]Department of Computer Science and Engineering, Narasaraopeta Engineering College, Narasaraopeta, Andhra Pradesh, India.
[5]Department of Computer Applications, Marian College Kuttikkanam Autonomous, Kerala, India.
[6]Symbiosis Institute of Computer Studies and Research (SICSR), Symbiosis International (Deemed University), Pune, Maharashtra, 411016, India.
[1]vetrigold@gmail.com, [2]shamik1980@gmail.com, [3]iamanshumehta@gmail.com, [4]gaddamsaranya4@gmail.com, [5]donamol.joseph@mariancollege.org, [6]abhinavgsits7@gmail.com

Correspondence should be addressed to Vetrithangam D : vetrigold@gmail.com

ArticleInfo
Journal of Machine and Computing (https://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc2025505134.
Received 07 March 2025; Revised from 29 April 2025; Accepted 16 June 2025.
Available online 05 July 2025.

Abstract – In modern industrial environments, early and accurate machine fault diagnosis is crucial for minimizing downtime, reducing maintenance costs, and ensuring operational safety. This research presents a robust fault classification framework that combines Recursive Feature Elimination with Cross-Validation (RFECV) and Random Forest classifiers to address the challenges of high dimensionality, overfitting, and limited model generalization. The proposed approach begins with comprehensive data preprocessing, followed by RFECV to identify and retain the most relevant features, thereby enhancing model efficiency and accuracy. Subsequently, a Random Forest classifier is trained on this optimized feature set to classify four fault types: No Failure, Power Failure, Tool Wear Failure, and Overstrain Failure. By integrating feature selection with ensemble learning, the framework effectively mitigates high variance and improves robustness under varying operational conditions and data distributions. Experimental results demonstrate that the proposed methodology achieves a high predictive accuracy of 99.2% along with improved computational efficiency, making it highly suitable for real-time fault diagnosis applications in smart manufacturing systems.

Keywords – Machine Fault Diagnosis, Random Forest, Recursive Feature Elimination, Feature Selection, Predictive Maintenance.

I. INTRODUCTION

In today's rapidly evolving industrial landscape, the integration of intelligent manufacturing systems has become a cornerstone for achieving operational excellence and competitive advantage. As industries increasingly embrace automation, the deployment of embedded sensors and condition-monitoring technologies has revolutionized how machines are monitored and maintained [1]. Predictive maintenance and fault diagnosis have emerged as essential components within this paradigm, enabling organizations to anticipate equipment failures before they occur, thereby minimizing downtime, reducing maintenance costs, and enhancing safety standards [2]. This shift from traditional reactive or scheduled maintenance to proactive and condition-based approaches relies heavily on advanced data-driven methods capable of extracting meaningful insights from vast amounts of sensor data. Using machine learning, it is now possible to analyze historical equipment data and identify complex patterns that change the way fault diagnosis takes place [3]. Machine learning techniques are widely appreciated for their simplicity, reliability, and fast training capabilities, making them suitable for diagnosing relatively simple systems. Deep Learning approaches, on the other hand, offer powerful end-to-end solutions capable of managing complex systems and compound faults, especially when large training datasets are available. Transfer learning methods address the critical issues of data scarcity and sample imbalance by enabling knowledge transfer across different operating conditions, machines, or even application domains. Despite these advancements, the

implementation of machine learning in real-world fault diagnosis continues to face challenges, particularly as engineering systems grow in complexity [4]. When using knowledge from the past and typical situations, ML models become very accurate in detecting small errors and predicting potential malfunctions [5]. The Extreme Random Forest (ERF) method was introduced to enhance feature extraction capabilities while reducing computational complexity. In this approach, high-dimensional data is projected into a lower-dimensional space using a randomly generated mapping matrix, effectively reducing dimensionality. This process not only lowers the computational burden but also improves classification performance after dimensionality reduction [6]. Predictive capabilities make it easier for manufacturing lines to change their maintenance processes and look after equipment more carefully. Even so, using fault diagnosis models in real factories is still very difficult. It is often necessary for those using older ML methods to have broad experience in the field and face problems with efficient computing [7][8]. Although deep learning is effective at dealing with difficult and complex data, it usually demands a lot of labelled samples, has severe computational needs, and remains unclear for users in terms of understanding how AI affects their operations.

In addition, industrial systems have many types of equipment, different working settings, and incomplete or noisy components. As a result of these factors, shifts in the data between how the model learns and its use in practice make the model perform poorly when it meets new or evolving errors. It is still an unsolved issue to maintain fault diagnostic models that are strong, expandable, and responsive on the spot even with these limitations [9]. Including a large number of instrument measurements in high-dimensional data can lead to redundant or irrelevant information, which may reduce the model's accuracy and increase computational requirements [10]. For this reason, Random Forests and other ensemble methods are used widely since they put multiple trees together, manage data that contains thousands of features, and give feature importance scores. Yet, Random Forest models may still be affected by the problem of too many variables and some of these may not matter for spotting faults. For this reason, using RFECV enables you to find the best subset of features step by step, removing unimportant ones as it goes, and constantly checks its effects on the model to prevent it from overfitting [11]. Random Forest (RF) is a robust ensemble learning method that constructs multiple decision trees and combines their outputs to improve classification accuracy and model stability compared to a single decision tree. Although numerous techniques have been explored for fault diagnosis, RF remains a valuable and necessary approach due to its fast execution speed, ability to handle high-dimensional data, and consistently strong performance in machinery fault diagnosis tasks [12]. Based on what this research learns, it suggests a strong machine fault diagnosis framework that uses Random Forest classification together with RFECV-based feature selection to enhance accuracy, make predictions clearer, and use resources more efficiently. In this system, we aim to separate four (No Failure, Power Failure, Tool Wear Failure, and Overstrain Failure) basic failure types that often come up in industrial machinery. First, it goes through careful preprocessing of the senses of vibration, torque, the time worked, and temperature. After that, the framework uses performance measures from cross-validation to help it remove unnecessary features as it progresses. Using the new tools in this feature set, the model is able to ensure accurate and prompt identification of faults.

Besides increasing the correctness of fault classification, the strategy also makes it easier to implement the model in real time since it cuts down on model complexity and processing power. That's why intelligent manufacturing settings are excellent places to use it due to the quick decisions and ability to withstand changes in its surroundings. Our aim through this research project is to give industry a solid, scalable, and clear method for diagnosing machine faults, so industrial operations become both safer and more efficient than before. The proposed methodology aims to implement Recursive Feature Elimination with Cross-Validation (RFECV) to effectively select the most significant features from the available dataset, which helps reduce dimensionality and enhances both the efficiency and accuracy of the fault diagnosis model. Building on this, a Random Forest classifier is developed and trained using the optimally selected features to accurately classify machine fault types, including No Failure, Power Failure, Tool Wear Failure, and Overstrain Failure, while addressing issues such as overfitting and improving model generalization. Furthermore, this approach tackles the challenges of high variance and limited robustness found in existing machine fault diagnosis methods by integrating feature selection with ensemble learning techniques, thereby ensuring reliable fault prediction across diverse operating conditions and varying data distributions.

II. LITERATURE REVIEW

Zhao et al. [13] proposed a novel framework named Identification for Fault Diagnosis (I4FD) that integrates regularized data-driven modeling and frequency analysis for machinery fault diagnosis under nonlinear system identification. The framework is designed to mitigate the effects of external environmental changes and improve diagnostic accuracy. It introduces a fault diagnosis-oriented regularization (FDoR) technique that incorporates prior physical knowledge through a penalty parameter, making the model specifically tailored for fault diagnosis applications. Unlike traditional approaches, I4FD supports continuous dynamic modeling using updated data. After model identification, frequency analysis is applied to extract fault-sensitive features. The framework achieves an accuracy of 92% on simulation and real-world cases. The advantage of I4FD is its ability to adapt to dynamic environments and deliver high accuracy, while a technical gap lies in the computational complexity and potential tuning challenges of the regularization process. Bode et al. [14] proposed a data-driven Fault Detection Algorithm (FDA) for heat pump systems, addressing the issue of reduced energy efficiency and potential system failures due to undetected faults in building heating and cooling systems. The model leverages big data approaches and AI techniques, using features extracted from a comprehensive fault dataset provided by the National

Institute of Standards and Technology (NIST). The FDA is trained on this lab-generated data and then applied to a real-world air-water heat pump system without system modifications. The model achieved an accuracy of 85% on the NIST dataset. The advantage of this approach lies in its cost-efficiency and use of detailed fault feature analysis from long-term monitoring data, which avoids the need for expensive custom setups. However, the model performs poorly on real-world data, highlighting a technical gap in generalizability due to domain shift, data incompleteness, and inadequate fault labeling in practical applications. Brito et al. [15] proposed a novel unsupervised framework for fault detection and diagnosis in rotating machinery, addressing the challenge of limited labeled data and the need for model interpretability. The approach consists of three main modules: feature extraction (from vibration signals in time and frequency domains), anomaly-based fault detection, and fault diagnosis using SHAP for model explainability.

To diagnose faults, the model leverages feature importance scores from SHAP explanations, enabling unsupervised classification and root cause analysis. The proposed methodology demonstrated its effectiveness on three rotating machinery datasets, achieving a maximum unsupervised classification accuracy of 96.72%, particularly with Ensemble, kNN, and CBLOF algorithms. The advantages of the proposed model include modularity in algorithm selection, interpretability using SHAP, and high accuracy without requiring labeled data. Nevertheless, the weak points in the area are that usefulness of the tool depends on the quality of the features, and methods such as SHAP and Local-DIFFI are computationally demanding.

Chen et al. [16] designed a machine learning model to detect and diagnose faults in real time in brushless motors, Support Vector Machines (SVM), Neural Networks (NN), and Random Forests are used (RF). It collects and combines information from numerous sensors to spot faults and check their degree of severity. offering ideas on how to counter the effects. Experiments prove that NN comes out on top in terms of success rate. SVM and RF performed very similarly, each having an accuracy of 95% and 92% respectively, while the best performance was given by NB with 97%. The main benefit of this method is that it improves the reliability, efficiency, and maintenance conditions. The use of brushless motors in industries. Still, there is a gap in technology when it comes to joining these the need to rapidly implement models in industries, considering they have to work continuously adapting to new types of faults as they happen. Tang et al. [17] proposed an intelligent fault detection system that uses DL for rotating speeds. Machinery that involves bearings, gears and gearboxes, and pumps. The framework tries to find ways to overcome the major problems linked to expert-dependent traditional faults diagnosis methods finding solutions by using only knowledge and manual work. With the help of deep learning, the framework lets users the automatic discovery of useful features and accurate recognition of types of faults. The model manages to reach an accuracy of 97.75%. It is an effective way to do extract features, since it reduces the amount of manual work. An intervention makes diagnostics more reliable and improves their consistency. However, it faces challenges in generalization, real-time application, and adaptability to unseen fault types, which are highlighted as areas for future research. Gonzalez-Jimenez et al. [18] proposed a machine learning-based fault diagnosis strategy for detecting power connection failures in induction machines, such as high resistance connections (HRC), single phasing faults, and opposite wiring connections. The model is designed to aid maintenance personnel in identifying these faults, particularly those caused by human errors during assembly. Due to the scarcity of real-world failure data, a simulation-driven approach using Software-in-the-Loop (SiL) simulations was adopted to generate synthetic training data. The proposed system achieved an accuracy of 98.5%. Using this approach, it's possible to identify a range of faults even without using real data. Its disadvantage is its dependence on simulations, which may decrease its effective use in real industries.

Tran et al. [19] proposed an IoT-based architecture integrated with machine learning algorithms to enhance cybersecurity in cyber-physical systems (CPS) for industrial electrical machines. The architecture focuses on monitoring induction motor status and detecting cyber-attacks in real time. The system uses the Random Forest algorithm for fault detection due to vibration and cyber-attack recognition, achieving an accuracy of 99.03%, which outperforms other ML models in industrial conditions. The infrastructure leverages the CONTACT Element IoT platform to visualize motor faults and fake data signals triggered by detected cyber-attacks on a dashboard. The advantage of this model lies in its high detection accuracy, low latency, and clear visualization, making it suitable for cost-effective and secure remote monitoring. However, technical gaps remain in terms of scalability across diverse industrial networks and robustness under varying attack types. Shubita et al. [20] proposed a machine learning-based fault diagnosis system that uses acoustic emission (AE) signals for early fault detection in rotating machines. The system is implemented on an embedded device with IoT connectivity, enabling real-time fault detection and classification. It achieved an accuracy of 96.1% using a fine decision tree model. The advantage of this approach is its ability to provide accurate and real-time monitoring with minimal latency, making it suitable for industrial deployment. However, the technical gap lies in the limited exploration of model robustness under varying operational or noisy conditions, which may affect real-world generalization.

Siyuan et al. [21] proposed a duplet classification model combining two 1-D Convolutional Neural Networks (CNNs) for fault diagnosis in rotating machinery involving both rotor and bearing components. The idea involved through the model was constructed by working on a dataset of 48 machine health problems created by different faults different levels and types of these two parties. CNN architecture has been created to distinguish between rotor and having the ability to respond to various external problems without getting damaged. It was possible to achieve the model. A high rate of identifying mixed faults at 95.93% proves that the results are highly reliable. Moreover, a single-vs-rest approach was built based on CNN information to catch known diseases. Four new fault categories, including those that go unnoticed, were tested by this study. Its usefulness comes from the fact that it is felt in many parts of society the ability to work in

complicated environments and recognize new types of faults. However, there is a technical challenge as using different models for each type of fault may increase the overall model. Real-time situations can cause major challenges due to lots of calculations involved. Shao et al. [22] introduced a fault diagnosis method that depends on deep learning (DBN) to detect the main status of induction motors by examining the distribution of their vibration signals. The model is made by putting several Restricted Boltzmann Machines (RBMs) on top of each other and training it in layers. It combines the steps of extracting features and doing the classification into one approach, so you do not have to engineer features manually. On data from the machine fault simulator, the accuracy of the classification is 99%.

Because this way works with raw information, the model can learn to structure the data and make the process of finding issues automatic and smart. Still, getting the right performance from the model requires careful selection of scale and depth, due to which tuning hyperparameters and running the model can be difficult.

Sohaib et al. [23] proposed a fault diagnosis method that combines a two-layer bearing with a hybrid set of data, along with SAEDNN. The model deals with finding patterns of faults and measurements of crack sizes from vibration signals that change with changing conditions and various fault levels in machines. It is more accurate than SVMs and BPNNs with an accuracy of 99.10%. Its real benefit is that it helps find more important features in the vibration data, making it easier to classify sounds under changing conditions. Kafeel et al. [24] proposed a fault detection method for rotating machines by studying the vibration signals. This system performs empirical mode decomposition (EMD) to filter noise from the signals and does multi-domain feature extraction to find both the time and frequency features of vibration data collected from healthy and bad induction motors. The extracted features are classified using multiple algorithms including SVM, KNN, Decision Tree, and Linear Discriminant Analysis, with the support vector machine using a Gaussian kernel achieving the best performance of 98.2% accuracy.

The advantage of this method lies in the hybrid use of time and frequency features, which enhances the fault discriminative capability of the model. However, a technical gap remains in the generalization of the system across different machine types and operational conditions, which could affect its applicability in broader industrial settings. Hung et al. [25] proposed a system-on-chip (SoC)-based tool wear detection model that leverages deep learning with sensor fusion techniques. The system was trained using vibrational and acoustic signals collected from a three-axis CNC machine operating under various spindle speeds and torque conditions. The inputs to the deep learning model were frequency spectrum representations of signals from a MEMS microphone and a three-axial accelerometer, with tool flank wear measured via a camera, adhering to ISO 8688-2:1989 standards. The model achieved detection accuracies of 99.7% for the single-sensor model and 87.75% for the fused model when deployed on a Pocket Beagle SoC.

The advantage of this system lies in its real-time detection capability, high accuracy, and cost-efficient embedded implementation. However, it shows reduced performance in the fused model, possibly due to signal integration complexity or variability in machining conditions, indicating a need for more robust fusion strategies. Orrù et al. [26] proposed a simple and easy-to-implement machine learning (ML) model for early fault prediction of centrifugal pumps in the oil and gas industry. The model is based on real-life sensor data including temperature, pressure, and vibration readings, which are pre-processed and denoised before training. Two algorithms—Support Vector Machine (SVM) and Multilayer Perceptron (MLP)—were implemented using the KNIME platform. The model achieved an accuracy of 98.1%, successfully detecting system deviations and issuing fault prediction alerts. The advantage of this approach lies in its practical simplicity and effective performance using real industrial data, supporting maintenance decision-making. However, the model is still in a preliminary stage, and potential technical gaps include the need for broader validation across different operating conditions and scalability for more complex fault scenarios.

Table 1. A Review of Research on Machine Fault Diagnosis Techniques

Author	Proposed Model	Findings	Challenges
Zhao et al. [13]	Identification for Fault Diagnosis (I4FD)	Achieved 92% accuracy in machinery fault diagnosis by integrating regularized NARX modeling and frequency analysis; incorporates physical knowledge via FDoR for continuous dynamic modeling.	Computational complexity and tuning difficulties in regularization parameters.
Bode et al. [14]	Data-driven Fault Detection Algorithm (FDA)	Achieved 85% accuracy in detecting faults in heat pump systems using AI-based FDAs trained on NIST laboratory data; enabled transfer to real-world systems without hardware modifications; leveraged big data and feature extraction for energy-efficient building climate systems.	Poor generalization to real-world data due to domain shift, incomplete data, and fault labeling issues.

Brito et al. [15]	Unsupervised Framework for Fault Detection and Diagnosis in Rotating Machinery	Achieved 96.72% accuracy in unsupervised classification using Ensemble, kNN, and CBLOF; employs SHAP-based explainability for root cause analysis; effective across three real-world rotating machinery datasets.	Computational cost of interpretability methods (e.g., SHAP, Local-DIFFI); performance sensitivity to the quality of extracted vibration features.
Chen et al. [16]	ML-based fault diagnosis using SVM, NN, and RF	Achieved 97% accuracy with NN, 95% with SVM, and 92% with RF; effectively analyzes fault severity and suggests countermeasures using sensor data.	Real-time integration challenges and limited adaptability to evolving fault patterns. The model faces a high variance issue as it struggles to validate on unseen faults
Tang et al. [17]	Deep Learning-Based Intelligent Fault Diagnosis Framework	Achieved 97.75% accuracy in fault classification for rotating machinery components (bearings, gears, pumps) by enabling automatic feature learning and reducing reliance on manual feature extraction.	Generalization issues, real-time implementation constraints, and difficulty adapting to unseen fault types. The model might lead to overfitting with increased epochs
Gonzalez-Jimenez et al. [18]	ML-Based Fault Diagnosis for Power Connections in IMs	Achieved 98.5% accuracy in diagnosing power connection faults (HRC, single phasing, and opposite wiring) using Software-in-the-Loop (SiL) simulation-generated training data.	Dependency on simulated data may limit real-world generalizability; lacks validation with field datasets.
Tran et al. [19]	IoT-based architecture with integrated ML (Random Forest) for CPS security and motor fault detection	Achieved 99.03% accuracy in detecting induction motor faults and cyber-attacks using Random Forest; leverages CONTACT Element IoT platform for real-time visualization of motor status and cyber-attack data; offers low latency, high detection accuracy, and clear dashboards.	Scalability across heterogeneous industrial networks and robustness under diverse attack scenarios remain open issues.
Shubita et al. [20]	ML-based Fault Diagnosis System using AE on IoT-Enabled Device	Achieved 96.1% accuracy in early fault detection of rotating machines using AE signals; implemented on embedded IoT device for real-time monitoring.	Limited robustness under varying operational/noisy conditions; lacks generalization to real-world environments.
Chen Siyuan et al.[21]	Duplet Classifier using two 1-D CNNs	Achieved 95.93% accuracy in diagnosing mixed faults in rotating machinery; utilizes two parallel CNNs to diagnose rotor and bearing faults separately; validated on 48 machine health conditions and four new fault types.	Increased model complexity due to separate CNNs; computational overhead during real-time deployment.
Shao et al. [22]	Deep Belief Network (DBN)-based Fault Diagnosis	Achieved 99% accuracy in fault diagnosis of induction motors by automatically learning features from vibration signal frequency distributions. Combines feature extraction and classification in a unified deep learning framework using stacked RBMs.	Model performance depends heavily on architecture scale and depth; introduces challenges in hyperparameter tuning and computational complexity.

Kafeel et al. [24]	Fault detection system based on Hybrid machine learning models	The hybrid use of time and frequency features, which enhances the fault discriminative capability of the model	Generalization of the system across different machine types and operational conditions
Hung et al. [25]	Deep learning with sensor fusion	This system provides real-time detection capability.	This model faces integration capability issues
Orrù et al. [26]	Support Vector Machine (SVM)	Detecting system deviations and issuing fault prediction alerts"	This model faces challenges in broader validation across different operating conditions and in scaling to more complex fault scenarios

As shown in **table 1**, the existing fault diagnosis models face several technical challenges, including high computational complexity and difficulties in tuning regularization and hyperparameters. Many models struggle with generalization issues, particularly when validating on unseen fault types or transferring from simulated or laboratory data to real-world scenarios, often due to domain shifts and incomplete or noisy data. Real-time implementation and integration remain problematic, especially for deep learning and ensemble methods with increased model complexity and computational overhead. Industry experts are also very concerned about the ability to scale these networks in many settings and how they will handle ever-changing threats. Furthermore, knowing how the AI model works is helpful, but it contributes to the model's complexity, and there are usually difficulties for models to maintain their results as faults evolve and work in more types of environments. This research addresses the technical gaps of high variance and overfitting commonly observed in machine fault diagnosis models, focusing on improving robustness and generalization in Random Forest-based predictive maintenance.

III. PROPOSED METHODOLOGY

This section describes the proposed methodology illustrated in the **Fig 1**, which presents a structured methodology for machine fault classification using a machine learning approach. The process begins with data preprocessing, which includes steps such as dropping irrelevant columns, label encoding of categorical data, feature and target separation, and finally, a train-test split to prepare the dataset for modelling. Following preprocessing, a feature selection technique is applied using Recursive Feature Elimination with Cross-Validation (RFECV) to identify and retain only the most significant features, thereby improving both model efficiency and accuracy. The selected feature set is then used to train a Random Forest Classifier, a robust ensemble learning algorithm known for its accuracy and resilience to overfitting. The classifier is trained to predict different types of equipment failures. For any new input instance, the model predicts one of the four possible outcomes: No Failure, Power Failure, Tool Wear Failure, or Overstrain Failure, thus enabling proactive maintenance and minimizing operational downtime.

Preprocessing

The initial phase prepares the raw input dataset for subsequent analysis.

Input Dataset

Let the raw input dataset be represented by Equation (1)

$$D = \{(x_1, x_1), (x_2, x_2), \dots, (x_N, x_N)\} \tag{1}$$

where x_i is a vector of features for the i^{th} instance, and y_i is its corresponding label. The dataset has N instances and M initial features.

Dropping Irrelevant Columns

As shown in equation (2). This step aims to remove features that do not contribute to the predictive power of the model. denote the set of irrelevant feature indices. After removing these columns, the dataset is transformed into a new feature set, as represented in Equation (2).

$$F_{\text{relevant}} = \{j \mid j \in / F_{\text{irrelevant}}\} \tag{2}$$

$$D' = \{(x_1', x_1), (x_2', x_2), \dots, (x_N', x_N)\} \tag{3}$$

where x_i' is x_i with columns in $F_{irrelevant}$ removed. As shown in Equation (3), the updated dataset D' consists of input-output pairs where each x_i' is derived from the original feature vector x_i by excluding the features indexed in $F_{irrelevant}$. This results in a reduced-dimensional representation that retains only the most relevant features for model training.

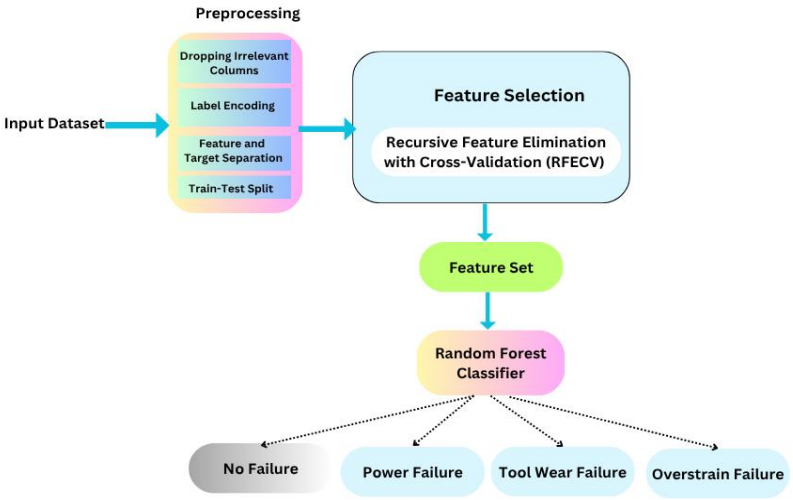

Fig 1. Architecture of the proposed model RFRFECV.

Label Encoding

If the target variable is categorical, it needs to be converted into numerical representations. Let $Y = \{y_1, y_2, y_3, \ldots, y_N$ (4) be the set of original categorical labels. As represented in equation (4) , the label encoding maps these to numerical values: L:Y→{0,1,2,3} (e.g., "No Failure" →0, "Power Failure" →1, etc.).

The transformed dataset now has numerical labels:

$$D'' = \{(x_1'', l_1), (x_2'', l_2), \ldots, (x_N'', l_N)\} \quad (4)$$

where li=L(yi).

Feature and Target Separation

The preprocessed dataset is split into features (X) and the target variable (y). $X=\{x_1'', x_2'', \ldots, x_N''\}$ (matrix of features) $y=\{l_1, l_2, \ldots, l_N\}$ (vector of target labels)

Train-Test Split

The dataset is divided into training and testing sets. Let D_{train}'' and D_{test}'' be the training and testing sets. $D_{train}''=(X_{train}, y_{train})$ $D_{test}''=(X_{test}, y_{test})$

Feature Selection

This stage identifies the most relevant subset of features.

Recursive Feature Elimination with Cross-Validation (RFECV)

RFECV recursively fits a model and removes the weakest features until the optimal number of features is reached based on cross-validation performance. Let M_{model} be the base machine learning model. Let K be the number of folds for cross-validation. The process can be described as follows:

Step 1: Initialization

Start with the full set of P features, $F=\{f_1, f_2, \ldots, f_P\}$.

Step 2: Iteration

The model is trained on the current feature set FFF using K-fold cross-validation applied to the training data X_{train}. During this process, the model's performance measured using metrics such as accuracy or F1-score—is evaluated on each fold. Let S_k represent the score obtained on fold k, and the average score across all folds is calculated as represented in equation (5).

$$\dot{s} = \frac{1}{K}\sum_{k=1}^{K} S_K \quad (5)$$

After evaluating performance, the feature with the lowest importance, denoted as $f_{weakest}$, is identified and removed from the feature set F. This iterative process continues to refine the model by eliminating the least significant features.

Step 3: Recursion

Repeat step 2 until an optimal performance is observed or a minimum number of features is reached.

Step 4: Optimal Feature Set Selection

Select the feature set $F_{selected}$ that yields the highest average cross-validation score. The dataset is then projected onto this selected feature set: $X_{train}'=X_{train}[F_{selected}]$ $X_{test}'=X_{test}[F_{selected}]$

Step 5: Feature Set

The output of the feature selection phase is the reduced set of features, $F_{selected}$.

Random Forest Classifier

The selected features are fed into a Random Forest Classifier for predicting the failure type.

Random Forest (RF): An ensemble learning method that constructs a multitude of decision trees.

Let T be the number of decision trees in the forest. Each tree $t \in \{1,\ldots,T\}$ is trained as follows:

Step 1 : Bootstrap Aggregating (Bagging)

A random subset of the training data X_{train}' (with replacement) is sampled to train each tree. Let this sample be $Dt' = (X_{train}, t', y_{train}, t)$.

Step 2: Random Feature Subspace

At each node of the decision tree, only a random subset of m features is considered for splitting.

Step 3: Tree Construction A decision tree T_t is grown on D_t'.

Step 4: Training

The Random Forest model, denoted as RF, is trained on the selected features of the training data: $RF = fit(X_{train}', y_{train})$

Step 5: Prediction

For a new, unseen instance x_{new} from X_{test}' (with features corresponding to $F_{selected}$), each tree t in the forest predicts a class $y\hat{}t$. The final prediction for x_{new} is the mode of the predictions from all trees: $y\hat{}^{new} = mode(y\hat{}^1, y\hat{}^2, \ldots, y\hat{}^T)$

Step 6 : Output Classes

The model outputs one of the four predefined failure types: "No Failure", "Power Failure", "Tool Wear Failure", "Overstrain Failure".

IV. RESULTS AND DISCUSSION

Dataset Description

The dataset used in this study contains detailed information related to engine performance and failure analysis. It includes variables such as vibration levels, torque, process temperature, air temperature (in Kelvin), engine speed (in RPM), and operational hours. Each entry is uniquely identified by a UDI (Unique Identifier) and is associated with a specific Product ID and engine type, where the type may denote categories such as motor (M) or liquid (L). The dataset also records the type of failure (if any), including specific classifications such as rotational failures, across a total of 500 machines. These attributes enable a comprehensive analysis of engine behavior under varying operational conditions. It can be used in many ways, for example, spotting reasons for engine failure, checking for engine temperature, speed, and torque, examining various engine types, and making forecasts for maintenance. The dataset is available at the following source link: https://www.kaggle.com/datasets/nair26/predictive-maintenance-of-machines. The dataset is split into 75% training and 25% testing.

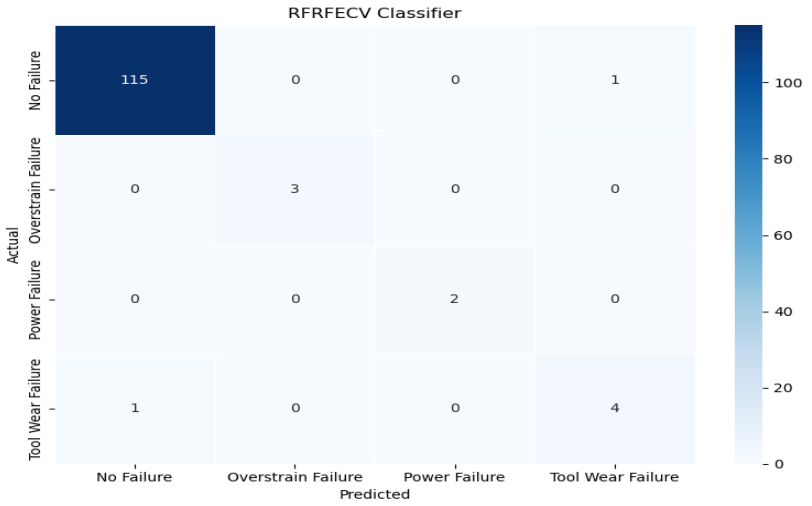

Fig 2. Confusion Matrix of The Proposed Model.

The **Fig 2** represents the prediction results of Random Forest classifier with Recursive Feature Elimination and Cross-Validation (RFRFECV) on multi-class machine failure problems. The four labels tested in the model were No Failure, Overstrain Failure, Power Failure, and Tool Wear Failure. It classified 115 instances as No Failure and just one was ruled as Tool Wear Failure. All three cases of Overstrain Failure were grouped under the correct class with 100% correctness. No errors happened in the prediction of Power Failure, as all two instances were accurately classified, and although four

instances of Tool Wear Failure were found, the model misclassified one as being from the No Failure class. On the whole, the confusion matrix confirm that the main class is classified very accurately and that all failure categories are detected well. The findings prove that choosing the right features and training the model correctly worked well. The slight number of cases that were wrongly classified implies that some failure groups may have traits in common with others. Therefore, RFRFECV was a dependable choice for handling data from many types of machinery and for recognizing faults in machines with preventive measures.

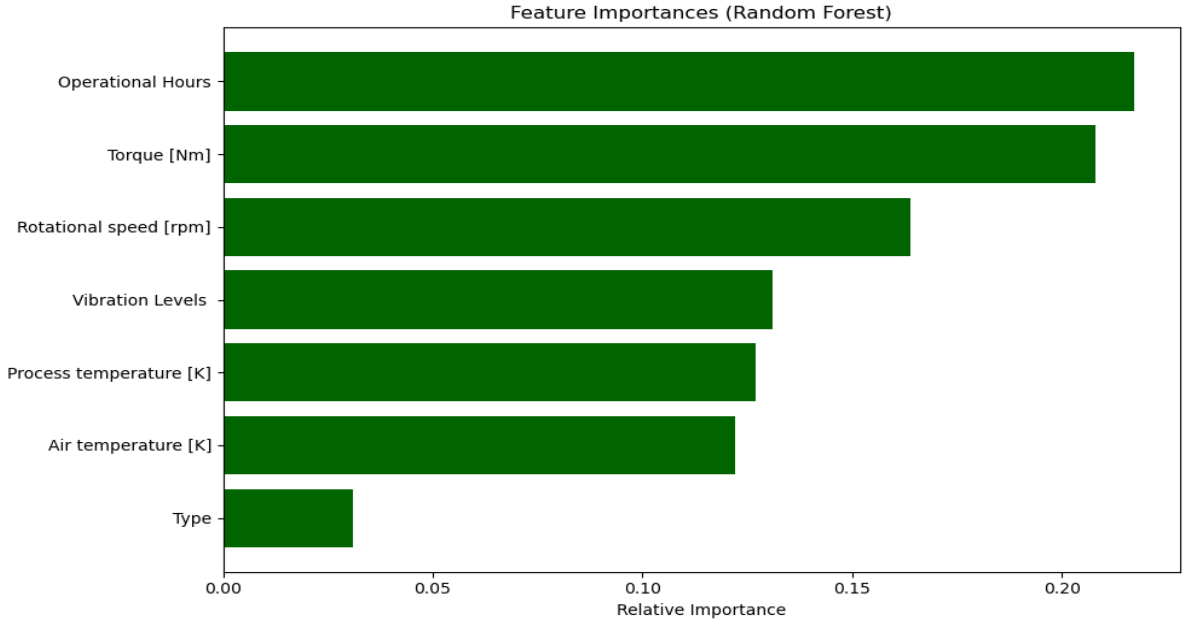

Fig 3. Feature Importance Analysis Using RFRFECV.

The **Fig 3** illustrates how a Random Forest classifier worked well when it was trained using RFRFECV to predict multiple machine failure conditions. There were four categories used in this classification problem: No Failure, Overstrain Failure, Power Failure, and Tool Wear Failure. This model was able to identify 115 of the instances in the "No Failure" category and just one case was wrongly marked as involving "Tool Wear Failure." When it comes to the "Overstrain Failure" category, the system did not make any mistakes and identified all the instances correctly. Thus, the model has the ability to tell between routine conditions and certain types of failures. Consequently, the RFRFECV method allowed the team to pick the right features, and this improved the model's precision in spotting and classifying different machine failures.

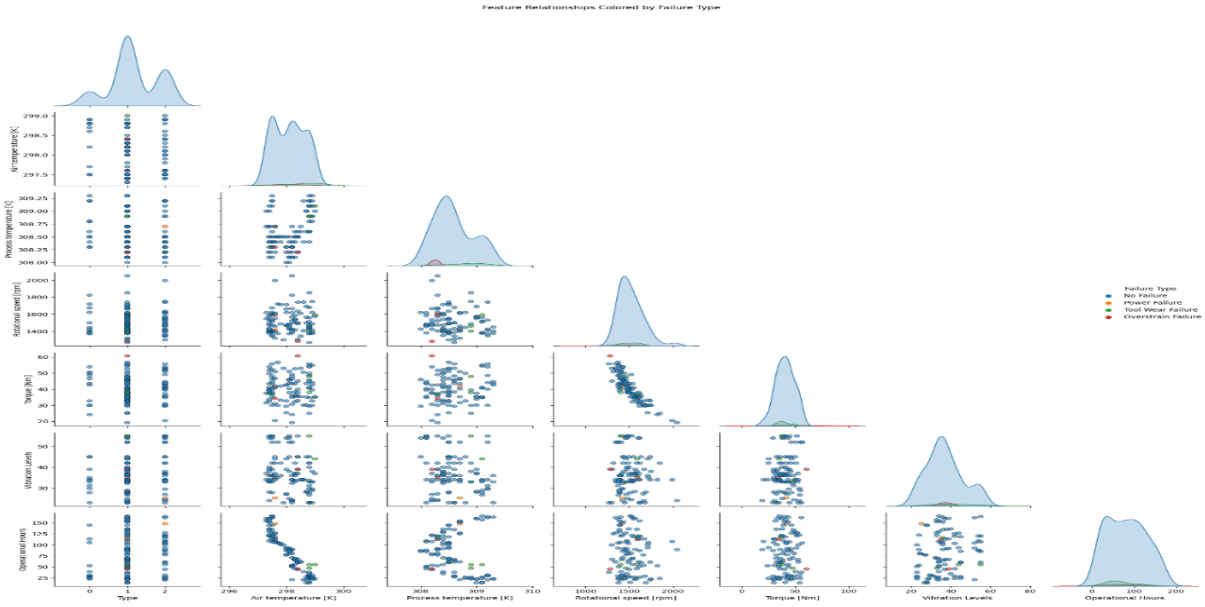

Fig 4. Feature Relationships of Failure Types in the Proposed Model.

The **Fig 4** presents Feature Relationships that are coloured by the type of failure that occurs. Scatter plots in the matrix display the connection between two variables in the data, and the diagonal plots indicate how each variable is spread. No

Failure is depicted as blue, Power Failure in orange, Tool Wear Failure in green, and Overstrain Failure in red on every graph. The way operational and environmental features are related, color-coded according to the 'Failure Type'. Seeing the results makes it possible to spot connections between various features and the different forms of failure. KDE estimates are drawn in the diagonal figures, offering views of the distributions of the features individually. Each scatter plot in the off-diagonal plots shows the trend between two features. The analysis using pair plots explains the features' distributions alone and their relationships with one another, as well as the significant patterns spotted for each failure case. The feature called 'Type' is a category, with 'Type 1' occurring most often, while both 'Air Temperature' and 'Process Temperature' are narrowly distributed and only take values inside certain ranges, but 'Air Temperature' sometimes drops below these ranges. 'Rotational Speed' displays multiple peaks, suggesting varied operating regimes, whereas 'Torque' and 'Vibration Levels' demonstrate unimodal distributions concentrated at lower values with a tail extending to higher levels. 'Operational Hours' presents a broader distribution, with a noticeable peak at lower values potentially indicating newer units or shorter operational cycles. In terms of bivariate relationships, a strong inverse correlation exists between 'Rotational Speed' and 'Torque', where increased rotational speed generally corresponds to decreased torque, a typical characteristic of mechanical systems with constant power output. No direct linear relationship is evident between 'Operational Hours' and either 'Torque' or 'Rotational Speed across the entire dataset, although specific failure types might exhibit localized clustering. 'Air' and 'Process temperatures' show an expected correlation with each other, but their relationships with other operational parameters like 'Torque' or 'Rotational Speed' are less pronounced linearly. Similarly, 'Vibration Levels' show scatter with other features, but no strong linear correlations are immediately apparent across the dataset. Crucially, the coloring by 'Failure Type' illuminates key patterns: 'No Failure' instances, representing the majority, are broadly distributed across all features, forming the primary clusters. 'Power Failure' instances are fewer and tend to cluster in specific regions, such as higher torque values at varying operational hours, or lower rotational speeds combined with higher torque, potentially indicating overload conditions. 'Tool Wear Failure' events are sparse but more prominent at higher operational hours, consistent with accumulated wear, and also appear at higher 'Vibration Levels', a common symptom of tool degradation. Finally, 'Overstrain Failure' events are very rare and consistently occur at extremely high 'Torque' values, aligning with the definition of overstrain.

The **Fig 5** visually illustrates the accuracy performance of each fault diagnosis model, with each bar uniquely colored to distinguish between different techniques. The RFRFECV Classifier, proposed in this study, achieves the highest accuracy of 99.20%, o[utperforming all other existing approaches. Notably, models such as the IoT-based architecture with integrated machine learning (99.03%), Deep Belief Network (99%), and Hybrid ML models (98.20%) also demonstrate strong performance, reflecting a clear trend toward the adoption of hybrid and deep learning-based solutions for fault diagnosis.

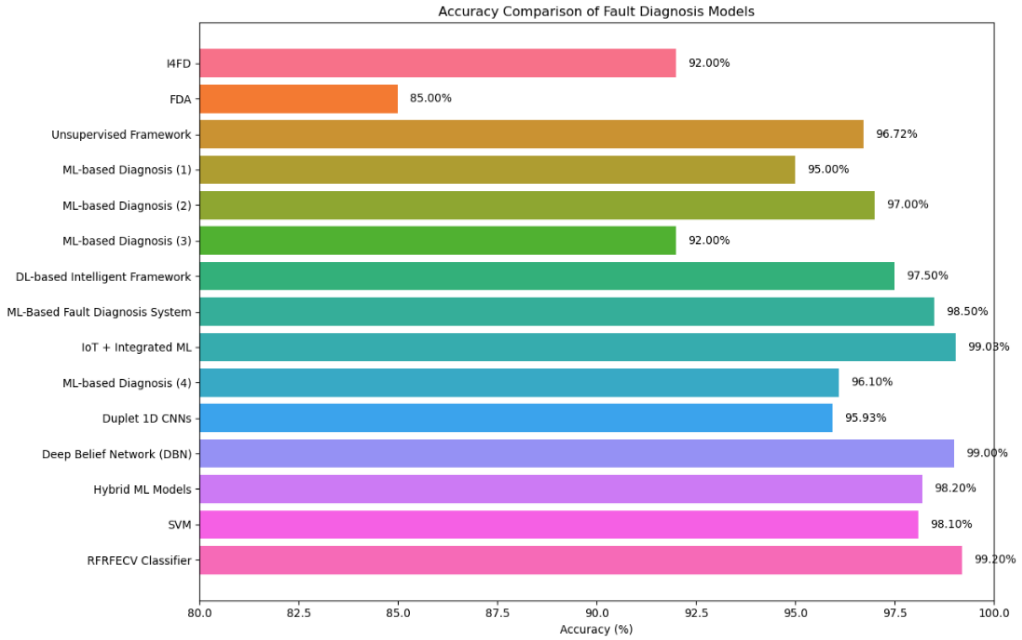

Fig 5. Accuracy Comparison of the Proposed Method.

V. CONCLUSION

This study presents a comprehensive machine fault diagnosis framework that effectively combines Recursive Feature Elimination with Cross-Validation (RFECV) and Random Forest classification to enhance predictive accuracy and model robustness. By systematically selecting the most significant features, the proposed approach reduces dimensionality, mitigates overfitting, and improves computational efficiency. The Random Forest classifier trained on the optimized feature set demonstrated exceptional performance, achieving an accuracy of 99.2% in classifying multiple fault types, including No Failure, Power Failure, Tool Wear Failure, and Overstrain Failure. This validates the effectiveness of integrating feature

selection with ensemble learning in addressing common challenges such as high variance and poor generalization. The framework's robustness and reliability make it well-suited for real-time fault diagnosis applications in smart manufacturing environments, ultimately contributing to improved operational safety and reduced maintenance costs. Future work may focus on extending this approach to other industrial domains and exploring adaptive methods to handle evolving fault patterns.

CRediT Author Statement

The authors confirm contribution to the paper as follows:

Conceptualization: Vetrithangam D, Shamik Palit, Anshu Mehta, Gaddam Saranya, Donamol Joseph and Abhinav Pathak; **Writing- Original Draft Preparation:** Vetrithangam D, Shamik Palit, Anshu Mehta, Gaddam Saranya, Donamol Joseph and Abhinav Pathak; **Visualization:** Gaddam Saranya, Donamol Joseph and Abhinav Pathak; **Investigation:** Vetrithangam D, Shamik Palit and Anshu Mehta; **Supervision:** Gaddam Saranya, Donamol Joseph and Abhinav Pathak; **Validation:** Donamol Joseph and Abhinav Pathak; **Writing- Reviewing and Editing:** Vetrithangam D, Shamik Palit, Anshu Mehta, Gaddam Saranya, Donamol Joseph and Abhinav Pathak; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

No data was used to support this study.

Conflicts of Interests

The authors declare no conflict of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

References

[1]. N. E. Sepulveda and J. Sinha, "Parameter Optimisation in the Vibration-Based Machine Learning Model for Accurate and Reliable Faults Diagnosis in Rotating Machines," Machines, vol. 8, no. 4, p. 66, Oct. 2020, doi: 10.3390/machines8040066.

[2]. Z. Li, Y. Wang, and K.-S. Wang, "Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario," Advances in Manufacturing, vol. 5, no. 4, pp. 377–387, Dec. 2017, doi: 10.1007/s40436-017-0203-8.

[3]. Z. Xiao, Z. Cheng, and Y. Li, "A Review of Fault Diagnosis Methods Based on Machine Learning Patterns," 2021 Global Reliability and Prognostics and Health Management (PHM-Nanjing), pp. 1–4, Oct. 2021, doi: 10.1109/phm-nanjing52125.2021.9612779.

[4]. J. Cen, Z. Yang, X. Liu, J. Xiong, and H. Chen, "A Review of Data-Driven Machinery Fault Diagnosis Using Machine Learning Algorithms," Journal of Vibration Engineering & Technologies, vol. 10, no. 7, pp. 2481–2507, Apr. 2022, doi: 10.1007/s42417-022-00498-9.

[5]. M. Fernandes, J. M. Corchado, and G. Marreiros, "Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review," Applied Intelligence, vol. 52, no. 12, pp. 14246–14280, Mar. 2022, doi: 10.1007/s10489-022-03344-3.

[6]. J. Luo, Y. Liu, S. Zhang, and J. Liang, "Extreme random forest method for machine fault classification," Measurement Science and Technology, vol. 32, no. 11, p. 114006, Jul. 2021, doi: 10.1088/1361-6501/ac14f5.

[7]. R. K. Patel and V. K. Giri, "Feature selection and classification of mechanical fault of an induction motor using random forest classifier," Perspectives in Science, vol. 8, pp. 334–337, Sep. 2016, doi: 10.1016/j.pisc.2016.04.068.

[8]. G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine Learning for Predictive Maintenance: A Multiple Classifier Approach," IEEE Transactions on Industrial Informatics, vol. 11, no. 3, pp. 812–820, Jun. 2015, doi: 10.1109/tii.2014.2349359.

[9]. G. Xu et al., "Data-Driven Fault Diagnostics and Prognostics for Predictive Maintenance: A Brief Overview," 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), Aug. 2019, doi: 10.1109/coase.2019.8843068.

[10]. J. J. Saucedo-Dorantes, M. Delgado-Prieto, R. A. Osornio-Rios, and R. de Jesus Romero-Troncoso, "Multifault Diagnosis Method Applied to an Electric Machine Based on High-Dimensional Feature Reduction," IEEE Transactions on Industry Applications, vol. 53, no. 3, pp. 3086–3097, May 2017, doi: 10.1109/tia.2016.2637307.

[11]. B.-S. Yang, X. Di, and T. Han, "Random forests classifier for machine fault diagnosis," Journal of Mechanical Science and Technology, vol. 22, no. 9, pp. 1716–1725, Sep. 2008, doi: 10.1007/s12206-008-0603-6.

[12]. Misra, P., & Yadav, A. S, Improving the classification accuracy using recursive feature elimination with cross-validation. Int. J. Emerg. Technol, 11(3), 659-665.2020.

[13]. Y. Zhao, Z. Liu, Z. Yang, Q. Han, and H. Ma, "Machinery fault diagnosis-oriented regularization for nonlinear system identification: Framework and applications," Applied Acoustics, vol. 231, p. 110537, Mar. 2025, doi: 10.1016/j.apacoust.2025.110537.

[14]. G. Bode, S. Thul, M. Baranski, and D. Müller, "Real-world application of machine-learning-based fault detection trained with experimental data," Energy, vol. 198, p. 117323, May 2020, doi: 10.1016/j.energy.2020.117323.

[15]. L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," Mechanical Systems and Signal Processing, vol. 163, p. 108105, Jan. 2022, doi: 10.1016/j.ymssp.2021.108105.

[16]. X. Chen, M. Wang, and H. Zhang, "Machine Learning-based Fault Prediction and Diagnosis of Brushless Motors," Engineering Advances, vol. 4, no. 3, pp. 130–142, Aug. 2024, doi: 10.26855/ea.2024.07.004.

[17]. S. Tang, S. Yuan, and Y. Zhu, "Deep Learning-Based Intelligent Fault Diagnosis Methods Toward Rotating Machinery," IEEE Access, vol. 8, pp. 9335–9346, 2020, doi: 10.1109/access.2019.2963092.

[18]. D. Gonzalez-Jimenez, J. del-Olmo, J. Poza, F. Garramiola, and I. Sarasola, "Machine Learning-Based Fault Detection and Diagnosis of Faulty Power Connections of Induction Machines," Energies, vol. 14, no. 16, p. 4886, Aug. 2021, doi: 10.3390/en14164886.

[19]. M.-Q. Tran, M. Elsisi, K. Mahmoud, M.-K. Liu, M. Lehtonen, and M. M. F. Darwish, "Experimental Setup for Online Fault Diagnosis of Induction Machines via Promising IoT and Machine Learning: Towards Industry 4.0 Empowerment," IEEE Access, vol. 9, pp. 115429–115441, 2021, doi: 10.1109/access.2021.3105297.

[20]. R. R. Shubita, A. S. Alsadeh, and I. M. Khater, "Fault Detection in Rotating Machinery Based on Sound Signal Using Edge Machine Learning," IEEE Access, vol. 11, pp. 6665–6672, 2023, doi: 10.1109/access.2023.3237074.

[21]. S. Chen, Y. Meng, H. Tang, Y. Tian, N. He, and C. Shao, "Robust Deep Learning-Based Diagnosis of Mixed Faults in Rotating Machinery," IEEE/ASME Transactions on Mechatronics, vol. 25, no. 5, pp. 2167–2176, Oct. 2020, doi: 10.1109/tmech.2020.3007441.

[22]. S.-Y. Shao, W.-J. Sun, R.-Q. Yan, P. Wang, and R. X. Gao, "A Deep Learning Approach for Fault Diagnosis of Induction Motors in Manufacturing," Chinese Journal of Mechanical Engineering, vol. 30, no. 6, pp. 1347–1356, Oct. 2017, doi: 10.1007/s10033-017-0189-y.

[23]. M. Sohaib, C.-H. Kim, and J.-M. Kim, "A Hybrid Feature Model and Deep-Learning-Based Bearing Fault Diagnosis," Sensors, vol. 17, no. 12, p. 2876, Dec. 2017, doi: 10.3390/s17122876.

[24]. Kafeel et al., "An Expert System for Rotating Machine Fault Detection Using Vibration Signal Analysis," Sensors, vol. 21, no. 22, p. 7587, Nov. 2021, doi: 10.3390/s21227587.

[25]. C.-W. Hung, C.-H. Lee, C.-C. Kuo, and S.-X. Zeng, "SoC-Based Early Failure Detection System Using Deep Learning for Tool Wear," IEEE Access, vol. 10, pp. 70491–70501, 2022, doi: 10.1109/access.2022.3187043.

[26]. P. F. Orrù, A. Zoccheddu, L. Sassu, C. Mattia, R. Cozza, and S. Arena, "Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry," Sustainability, vol. 12, no. 11, p. 4776, Jun. 2020, doi: 10.3390/su12114776.