Journal of Dynamics, Monitoring and Diagnostics, 2025, 4, 160-169
https://doi.org/10.37965/jdmd.2025.771

ISTP
RESEARCH ARTICLE

Towards Fault Diagnosis Interpretability: Gradient Boosting Framework for Vibration-Based Detection of Experimental Gear Failures

Auday Shaker Hadi and Luttfi A. Al-Haddad
Mechanical Engineering Department, University of Technology - Iraq, Baghdad, Iraq

(Received 31 March 2025; Revised 30 July 2025; Accepted 13 August 2025; Published online 13 August 2025)

Abstract: Accurate and interpretable fault diagnosis in industrial gear systems is essential for ensuring safety, reliability, and predictive maintenance. This study presents an intelligent diagnostic framework utilizing Gradient Boosting (GB) for fault detection in gear systems, applied to the Aalto Gear Fault Dataset, which features a wide range of synthetic and realistic gear failure modes under varied operating conditions. The dataset was preprocessed and analyzed using an ensemble GB classifier, yielding high performance across multiple metrics: accuracy of 96.77%, precision of 95.44%, recall of 97.11%, and an F1-score of 96.22%. To enhance trust in model predictions, the study integrates an explainable AI (XAI) framework using SHAP (SHapley Additive exPlanations) to visualize feature contributions and support diagnostic transparency. A flowchart-based architecture is proposed to guide real-world deployment of interpretable fault detection pipelines. The results demonstrate the feasibility of combining predictive performance with interpretability, offering a robust approach for condition monitoring in safety-critical systems.

Keywords: explainable AI; gears; Gradient Boosting; vibration signals

I. INTRODUCTION

A. GENERAL BACKGROUND

The reliability and safety of mechanical systems are paramount in modern industrial environments [1–3]. As machines become increasingly sophisticated and integrated into core operations, the ability to diagnose faults in advance and with precision is now a key consideration in preventing surprise failures to reduce maintenance costs, and ensuring uninterrupted operation [4–7]. Traditional fault-diagnostic methods rely on human inspections, threshold-based alarms, or model-based methods with intensive domain knowledge and assumptions on simplified systems. Though such methods have served in various applications, they may not be able to handle dynamic operating conditions and complex patterns of faults.

In recent years, the emergence of intelligent systems, particularly those based on machine learning (ML) and artificial intelligence (AI), has brought transformative changes to the field of fault diagnosis [8,9]. These data-driven techniques can learn from historical and real-time sensor data, identify subtle patterns associated with different failure modes, and provide accurate classification or prediction of system health states [10,11]. Such methods offer the advantage of adaptability and scalability, especially when applied to large-scale or condition-driven monitoring systems.

However, despite their outstanding performance, the deployment of intelligent diagnostic systems into key industries remains cautious. The reasons behind this are primarily interpretability, trustworthiness, and the transparency of decision-making. Advanced AI models are seen as "black boxes" making predictions with no visible reasoning, and it is difficult for engineers and operators to verify, trust, and act on the predictions. Consequently, this issue has brought about increased focus on the domain of explainable AI (XAI), whose aim is to make AI models interpretable and transparent without diminishing precision. As intelligent monitoring systems evolve, the inclusion of accurate and interpretable diagnostic tools is a technological imperative and practical necessity for building trust and accountability for automated fault diagnosis.

Among the various ML algorithms used in fault diagnosis, Gradient Boosting (GB) stands out for its high predictive accuracy and its ability to handle complex nonlinear relationships [11]. GB operates by building an ensemble of weak learners of decision trees in a stage-wise manner to iteratively minimize prediction error. While GB is not inherently interpretable like linear models, its decision-tree foundation makes it highly compatible with post-hoc explainability techniques such as SHAP (SHapley Additive exPlanations). SHAP assigns contribution scores to each feature for individual predictions, which uncovers the model's decision process in a way that is intuitive and quantifiable. This synergy between GB's learning capability and SHAP's explainability is particularly valuable in industrial diagnostic applications where both accuracy and transparency are critical.

1. LITERATURE REVIEW. In recent years, the field of intelligent fault diagnosis has seen a surge in research contributions leveraging ML and deep learning techniques to enhance the detection, classification, and prediction of mechanical failures. As shown in Table I, various studies have adopted a wide range of approaches, including deep neural networks, support vector machines, ensemble methods, and hybrid models that fuse physical and data-driven methodologies. These techniques have been applied across

Corresponding author: Luttfi A. Al-Haddad (e-mail: Luttfi.a.alhaddad@uotechnology.edu.iq).

numerous mechanical systems such as gearboxes, bearings, and rotors.

Some of the works discussed here center on the analysis of the vibration signal due to its susceptibility to mechanical anomalies. Despite the ability of deep learning methods to achieve high diagnostic accuracy, the resulting models lack interpretability. Conventional ML models, on the other hand, present better interpretability but are feature-engineering intensive and less robust under complex fault conditions or imbalanced data. Some of the works also explore the use of hybrid or fusion models for overcoming these limitations and achieving robustness under varying fault conditions and severities. Even with these advances, the trade-off of accuracy with explainability is still an active research topic, especially for safety-critical applications where model decision trust is of utmost importance.

To complement the focused review in Table I, which centers on gear fault diagnosis approaches, Table II summarizes recent efforts in developing interpretable and explainable models for fault diagnosis and condition monitoring. These works reflect the current emphasis on transparency, trust, and human-centric AI systems in industrial applications throughout the utilization of tools such as SHAP, LIME, PDP, and Grad-CAM. This broader review positions the current work within the context of explainable diagnostic frameworks and highlights the novelty of applying GB in conjunction with SHAP for vibration-based gear fault detection.

While the methods listed in Table II represent a broad spectrum of XAI techniques, many of them face notable limitations when applied to vibration-based fault diagnosis. Techniques like LIME and KernelSHAP often rely on approximation methods that can introduce instability or inconsistency in feature attributions. Others, such as Grad-CAM or Layer-wise Relevance Propagation, are primarily designed for image or spatial data, which makes them less suitable for low-dimensional time-series signals. In addition, complex frameworks that incorporate knowledge graphs or hybrid reasoning systems often suffer from high computational overhead and limited generalizability across domains. These limitations highlight the need for simpler, more robust XAI solutions like the one proposed in this study, which balances diagnostic accuracy with transparent, domain-relevant feature contributions.

B. STUDY CONTRIBUTIONS

Despite the significant advancements in intelligent fault diagnosis, a major challenge persists in balancing high diagnostic accuracy with interpretability. Many state-of-the-art deep learning methods, while effective in complex fault classification, operate as black boxes and lack transparency, making them less suitable for safety-critical environments where explainable decision-making is essential. Additionally, some studies rely on limited or narrowly scoped datasets, which restrict the generalizability of the developed models across real-world fault scenarios.

To address this, the study proposes interpretable fault diagnosis using the GB, which offers a strong compromise between performance and explainability. By leveraging the comprehensive Aalto Gear Fault Dataset—rich in fault types, severities, and operational conditions—the framework is trained to deliver accurate predictions while maintaining insight into the model's decision-making process. Furthermore, a novel XAI architecture is introduced as it integrates SHAP analysis for transparent fault

Table I. Summary of recent studies in intelligent fault diagnosis

Ref.	ML Method	Application	Limitation
[25]	Deep Learning Neural Networks	Bearing fault detection	Needs large data
[26]	Multi-source Heterogeneous Data Fusion	Gearbox fault fusion	Complex architecture
[27]	Large Language Model-Based Deep Model	Composite fault diagnosis	High computation
[28]	Signal Processing + ML	Rotor-bearing classification	Manual feature design
[29]	Discrete Wavelet Transform + K-Star Algorithm	Gear fault classification	Low scalability
[30]	Feature Selection + Classification Algorithms	Fault severity estimation	Limited generalization
[31]	Virtual Physical + Data-Driven Fusion Model	Simulated-real fault fusion	Model tuning required
[32]	Design of Experiments + SVM	Spur gear detection	Parameter sensitivity
[33]	Naïve Bayes Classifier	Helical gear diagnosis	Assumes independence
[34]	Logistic Regression (SGD)	Helical gear monitoring	Linear assumptions

Table II. State-of-the-art studies on XAI techniques for fault diagnosis in industrial systems

Ref.	XAI/ML Method	Application/Highlight
[35]	SHAP, Grad-CAM, Deep Taylor, Smooth Simple Taylor	XAI-based bolt-loosening detection using 1D CNN and Lamb waves
[36]	SHAP, Local-DIFFI	Unsupervised fault detection and diagnosis in rotating machinery
[37]	LIME, SHAP, PDP, ICE	Predictive maintenance of rotating machines with multiple AI models
[38]	LLM + Knowledge Graphs (To-FD-EKG)	Interpretable reasoning in LLM-based fault diagnosis scenarios
[39]	Layer-wise Relevance Propagation	CNN explainability for guided wave damage detection in structures
[40]	LIME, SHAP	Interpretable motor sound classification using ANN, SVM, KNN, RF
[41]	LIME, SHAP	Interpretable fault diagnosis in aircraft landing gear using a two-tier ML framework; enhances trust, maintenance decisions, and safety

interpretation, thereby bridging the gap between diagnostic accuracy and model explainability. Unlike other XAI methods listed in Table II, which are often designed for high-dimensional data or localized interpretability, the current approach leverages GB with SHAP to offer both high diagnostic accuracy and globally interpretable feature attributions tailored for low-dimensional vibration data.

The remainder of this paper is structured as follows: Section II introduces the ML methodology with a focus on GB. Section III outlines the experimental dataset. Section IV presents and discusses the diagnostic results. Finally, Section V concludes the study.

II. MACHINE LEARNING: METHODOLOGY

A. MACHINE LEARNING

ML has become a fundamental approach in predictive analytics and intelligent decision-making, particularly in condition monitoring and fault diagnosis [12–14]. ML algorithms can uncover hidden patterns from data, learn complex relationships between variables, and make accurate predictions with minimal human intervention. Among various ML methods, ensemble learning techniques—those that combine multiple models—have gained popularity due to their robustness and enhanced performance. One such method is GB, which constructs a strong predictive model by sequentially combining weaker learners, typically decision trees, to minimize the prediction error through gradient descent optimization [15–18].

The GB algorithm works by minimizing a chosen loss function through a stage-wise process. Initially, the model $f_0(x)$ is set by finding a constant value that minimizes the overall loss, as shown in equation (1). In each iteration of m, the algorithm computes the pseudo-residuals, which represent the negative gradient of the loss function with respect to the current prediction—this is expressed in equation (2). These residuals guide the training of the next weak learner (typically a decision tree). Once the tree is trained, the optimal leaf output value γm is calculated by minimizing the loss within each leaf, as defined in equation (3).

$$f_o(x) = \underset{\gamma}{argmin} \sum_{i=1}^{N} L(yi,\gamma) \tag{1}$$

$$r_{im} = \left.\frac{-\partial L(yi,f(xi))}{\partial f(xi)}\right|_{f=fm-1}, \quad i = 1,2,\ldots,N \tag{2}$$

$$\gamma_m = \underset{\gamma}{argmin} \sum_{i\in leaf} L(yi, fm-1(xi) + \gamma) \tag{3}$$

The model is then updated by adding the scaled prediction of the new learner to the existing model, as indicated in equation (4). Finally, the complete model after M iterations is the sum of all previous learners' contributions, weighted by the learning rate ν, as shown in equation (5). This iterative process ensures that each new learner focuses on correcting the errors made by the previous ones, gradually improving prediction accuracy.

Table III. GB optimal parameters

Parameter	Value
Learning Rate (ν)	0.1
No. of Estimators	200
Maximum Depth	5
Subsample	0.8
Loss Function	Mean Squared Error
Split Criterion	Friedman MSE

$$f_m(x) = f_{m-1}(x) + \nu \cdot \gamma_m h_m(x) \tag{4}$$

$$f_M(x) = f_o0(x) + \nu \sum_{m=1}^{M} \gamma_m h_m(x) \tag{5}$$

In the equations above, $fm(x)$ represents the model at iteration m, ν is the learning rate controlling the contribution of each weak learner, $hm(x)$ is the weak learner (usually a decision tree), L is the loss function, and γm is the optimal update value. These parameters are central to the convergence and accuracy of the GB process. The selected hyperparameters for this study are summarized in Table III.

GB hyperparameters were tuned using a reproducible, two-stage procedure. The dataset was stratified into train/validation/test splits (70/15/15) with a fixed seed to prevent leakage across windows. On the training set, stratified 5-fold cross-validation was performed, with macro F1-score used as the optimization metric. A randomized search was first applied to explore a broad space—learning_rate {0.01–0.20}, n_estimators {100–500}, max_depth {3–6}, subsample {0.6–1.0}, min_samples_split {2–10}, min_samples_leaf {1–4}, max_features {"sqrt," None}—followed by a focused grid search around the best-performing region. Where supported, validation-based early stopping (n_iter_no_change=20, tol=1e-4) was employed. The optimal configuration (learning_rate=0.10, n_estimators=200, max_depth=5, subsample=0.80, min_samples_split=2, min_samples_leaf=1, max_features=None) was then retrained on the combined train and validation sets and evaluated once on the held-out test set. The hyperparameter tuning process is outlined in Fig. 1.

This minimalist feature set supports the goal of maintaining model transparency when coupled with SHAP-based interpretability, and avoids the complexity and redundancy often associated with high-dimensional time- or frequency-domain feature sets. Moreover, prior studies have validated the sensitivity of these statistical descriptors in rotating machinery health monitoring [10]. The schematic diagram of the methodology is depicted in Fig. 2 below.

B. EXPLAINABLE AI FRAMEWORK

To enhance trust, transparency, and operational integration in intelligent fault diagnosis, an XAI strategy is proposed in this study to complement the high-performing GB model. This study leverages SHAP's TreeSHAP algorithm, which is specifically optimized for tree-based models like GB, allowing for fast, exact, and consistent feature attributions compared to approximation-based SHAP methods used with other ML algorithms.

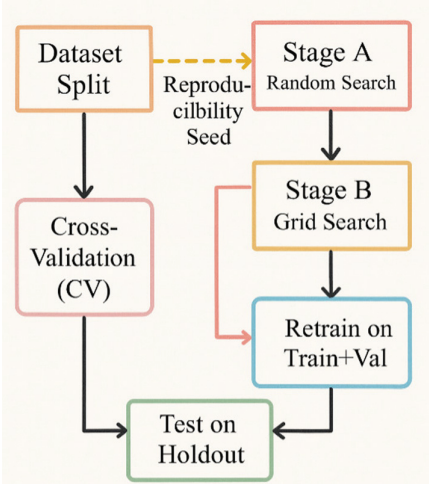

Fig. 1. Hyperparameter tuning workflow for the Gradient Boosting model.

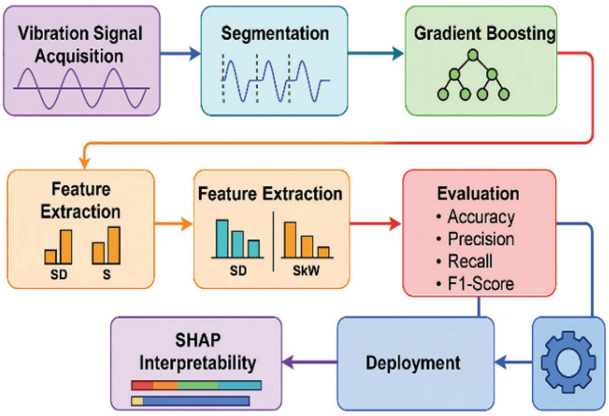

Fig. 2. Proposed methodology diagram.

As illustrated in Fig. 3, the framework begins with vibration signal acquisition and segmentation into 1000-sample intervals, followed by the extraction of key statistical features from multiple accelerometers. These features are then used to train and evaluate a GB classifier. To transition from a "black-box" to an interpretable system, SHAP (SHapley Additive exPlanations) is applied to quantify the impact of each feature on model predictions [19–21]. A decision checkpoint evaluates whether the SHAP-based explanations are interpretable to domain experts; if not, the process loops back to refine feature engineering or model complexity. Once interpretability is ensured, the model is deployed along with its explanation interface for real-time gear condition monitoring, which enables both accurate and transparent decision-making in safety-critical environments.

III. EXPERIMENTAL WORK

To develop and validate an interpretable fault diagnosis model, this study employed the Aalto Gear Fault Dataset (AGFD) [22]. The dataset is a comprehensive dataset of the vibrational signals measured from a reduced-scale azimuth thruster test bench under laboratory-controlled conditions.

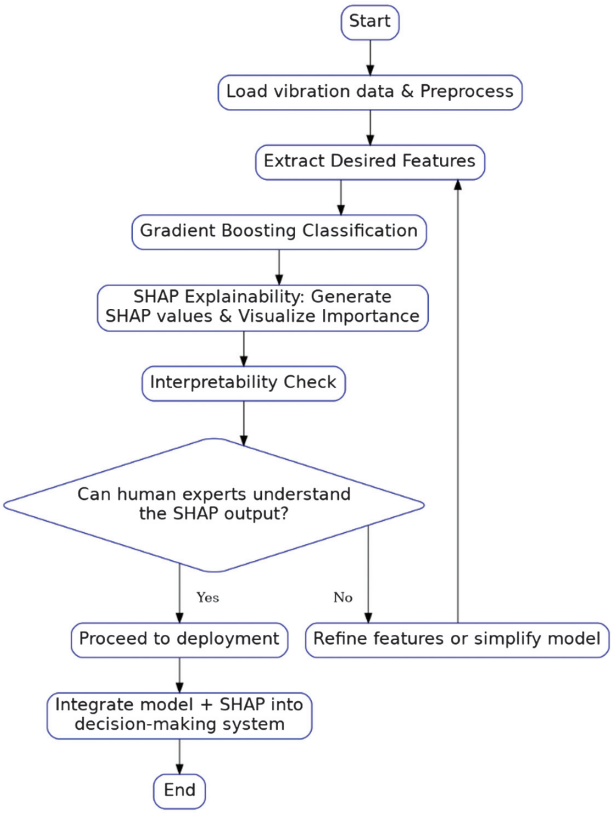

Fig. 3. Proposed XAI framework for Interpretable Gear Fault Diagnosis using GB and SHAP.

The dataset is especially designed for the evaluation of the performance of ML models for gear fault detection and classification because it contains synthetic-realistic cases of faults. The experimental platform used to generate the dataset is a scaled maritime thruster test bench designed to simulate the operational behavior of a real azimuthing thruster. The test bench incorporates two Bosch Rexroth synchronous servomotors—one acting as the driving motor and the other as a load simulator—connected through a drivetrain composed of multiple gearboxes and shafts. The mechanical layout includes 90-degree gearboxes with gear ratios of 3:1 and 4:1, elastomer couplings, flywheels, and a planetary gearbox, all configured to replicate torque transfer and dynamic loading observed in real-world marine propulsion systems. The system is instrumented with a total of 11 sensors, of which four accelerometers (A1–A4) were used in this study to capture the vibrational behavior of the gear system. These sensors were strategically placed at different locations on the drivetrain to ensure the collection of vibration signatures under various operating conditions and faults. Figure 4 illustrates the full experimental setup, while Fig. 5 shows the sensor layout and component arrangement of the thruster system.

The dataset includes two subgroups: the Aalto Shim Dataset (ASD), which contains synthetic faults created by attaching metal shims to gear teeth, and the AGFD, which contains replicated real-world failures such as tooth flank fracture, pitting, micropitting, and abrasive wear. For the purposes of this study, only the vibration signals from the accelerometers were used as input features. Furthermore, data were filtered to include only readings taken at a fixed rotational speed of 500 RPM that ensures consistent

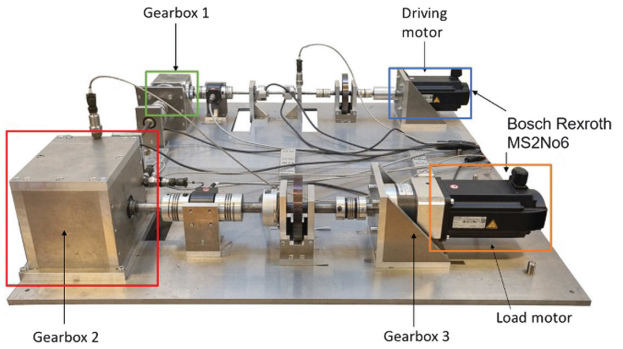

Fig. 4. Small-scale maritime thruster test bench of azimuth drive configuration [22].

operating conditions across all gear states. The collected time-series vibration data represents various mechanical states—healthy and faulty—under identical loading conditions, which isolate fault characteristics without interference from torque variability.

Each measurement file is organized into 12 columns. However, for this study, only columns 7–10, which represent the four accelerometer signals, were utilized. These columns contain vibration data in m/s^2 collected from sensors placed at key drivetrain locations. The current study utilizes the acceleration signals alone, measured at m/s^2 and captured using the Hansford HS-100 series sensor. The dataset includes various gear fault types, each replicated under two severity levels (mild and severe) using physical techniques designed to mimic real-world failure progression. Synthetic faults in the ASD dataset were created by attaching metal shims to gear teeth, while realistic faults in the AGFD dataset were introduced through procedures such as electrical discharge machining, abrasive blasting, and improper lubrication. The healthy state was also recorded for each gear pair and operating condition. The different fault types and the corresponding replication methods are summarized in Table IV.

This study used only the ASD dataset, which consists of 10 total classes—including 1 healthy class and 9 synthetic fault conditions. The operating speed was fixed at 500 RPM, and all conditions were recorded using a uniform test protocol to ensure consistency. These synthetic faults differ in shim thickness (0.01 mm, 0.03 mm, and 0.05 mm) and quantity (1–3), leading to subtle yet detectable variations in

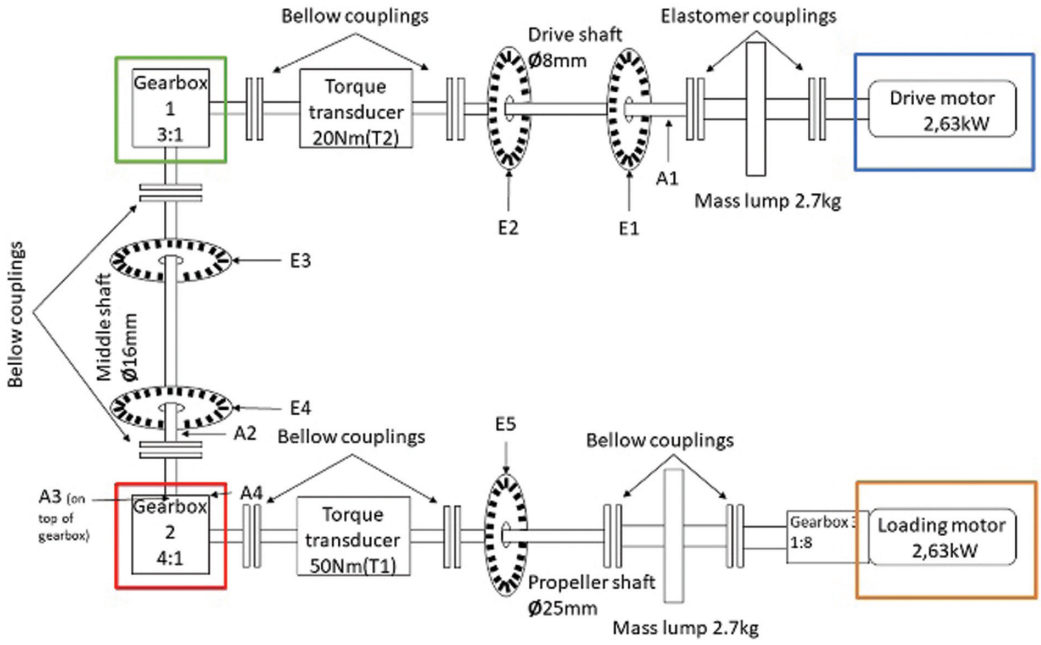

Fig. 5. Azimuth thruster model topology with accelerometer placements A1–A4 [22].

Table IV. Summary of gear fault types and replication methods

Fault type	Severity levels	Replication method	Description
Tooth Flank Fracture	Mild, Severe	Electrical Discharge Machining	Created incisions and complete tooth cuts at 45° angles to mimic cracking
Pitting/Spalling	Mild, Severe	Abrasive grinding	Small pits and extended surface removal to simulate damage progression
Micropitting	Mild, Severe	Sand blasting	Light to heavy frosting on gear surface representing material erosion
Abrasive Wear	Mild, Severe	Unlubricated long-duration operation	Progressive surface scoring and material loss across all gear teeth
Synthetic Shim Faults (ASD)	9 Fault Classes	Thin metal shims glued to gear teeth	Varied in thickness and count to simulate minor gear mesh anomalies

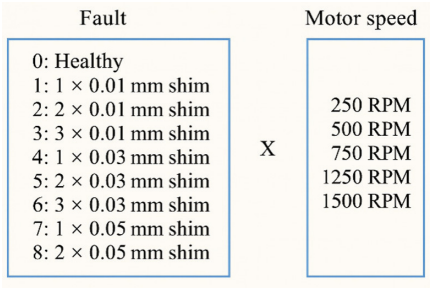

Fig. 6. Fault classes.

the vibration signatures. Figure 6 shows the classes and the parameters of the adopted method.

To evaluate the performance of the GB model, multiple standard error metrics were used. The root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and coefficient of variation of RMSE (CVRMSE) were computed using the following formulas with equations (6–9) as described below:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{6}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2} \tag{8}$$

$$CVRMSE = \frac{RMSE}{\bar{x}} \times 100 \tag{9}$$

These measurements provide a quantitative appreciation of the model's fit with the data. RMSE and MAE capture the magnitude of the error of prediction on average, and R^2 measures the proportion of variance explained by the model. CVRMSE normalizes the RMSE relative to the average, creating a scale-independent performance measure. Overall, this section introduced the GB algorithm as an interpretable method. It described the most significant training equations and performance metrics and enumerated the key tuning parameters, setting the stage for the experimental application of the next section.

To prepare the features for interpretable fault classification, standard deviation (SD) and skewness (S) were computed for each 1000-sample interval of the vibration signals. This interval-based statistical profiling enables a more structured and condensed representation of the raw vibration data. SD captures the degree of variability in the signal, which is often elevated in faulty gear conditions due to mechanical irregularities, impacts, or looseness [23]. Skewness reflects the asymmetry of the signal distribution, which may arise from shifts in gear wear [24]. These two features were selected based on their diagnostic relevance and ease of interpretation, as they can effectively distinguish between healthy and faulty signal patterns while keeping the feature space minimal [24]. The selection of SD and skewness was made prior to model training based on their diagnostic relevance in vibration analysis to capture variability and asymmetry in the signal, respectively; this domain-driven selection ensures simplicity throughout the diagnostic pipeline.

IV. RESULTS AND DISCUSSION

The time-domain vibration signals for all four accelerometers are shown in Fig. 7 through Fig. 10. Figure 7 illustrates the vibration patterns of acc1, with signal values ranging from a minimum of –7.97 m/s^2 to a maximum of 7.38 m/s^2 across all ten gear conditions. In Fig. 8, acc2 displays the highest range of vibration amplitudes, spanning from –15.40 m/s^2 to 13.77 m/s^2, indicating its sensitivity to fault-induced disturbances. Figure 9 presents the signal

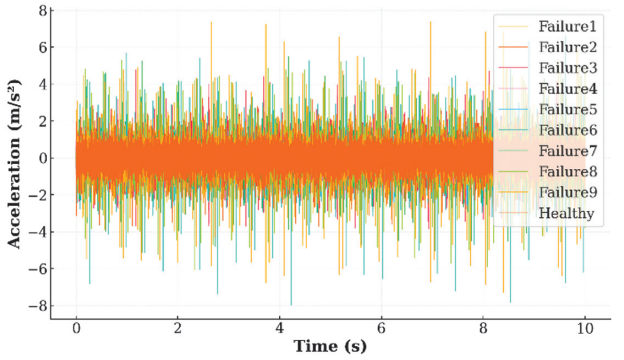

Fig. 7. Ten classes vibration signals of acc1.

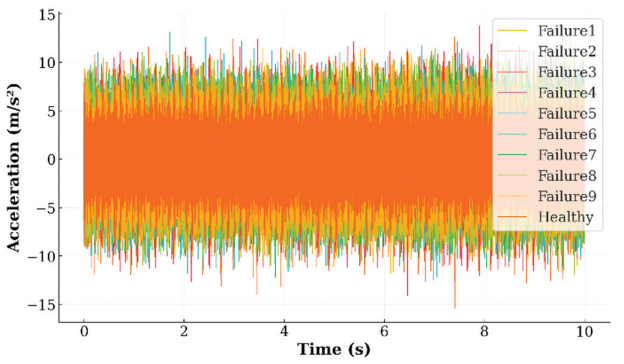

Fig. 8. Ten classes vibration signals of acc2.

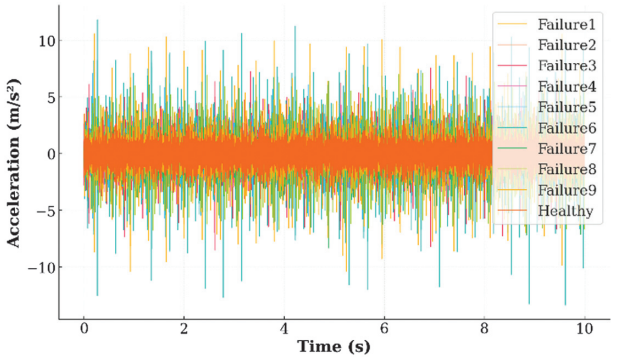

Fig. 9. Ten classes vibration signals of acc3.

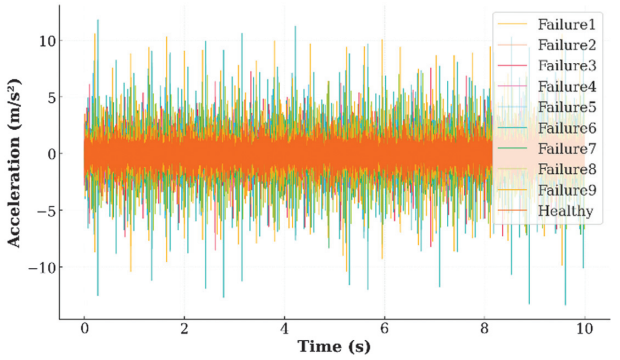

Fig. 10. Ten classes vibration signals of acc4.

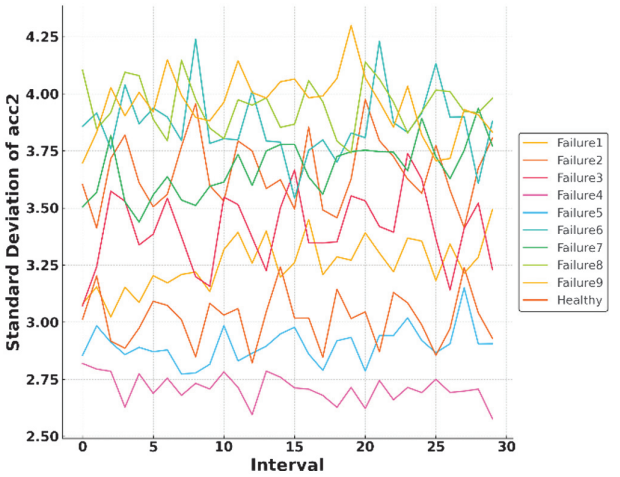

Fig. 11. Ten classes SD of acc2.

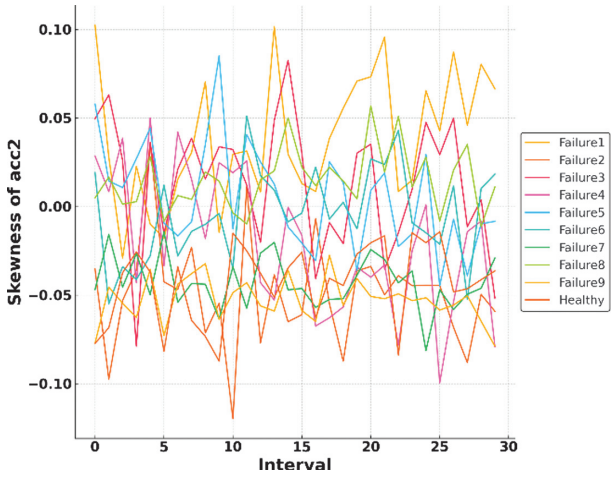

Fig. 12. Ten classes S of acc2.

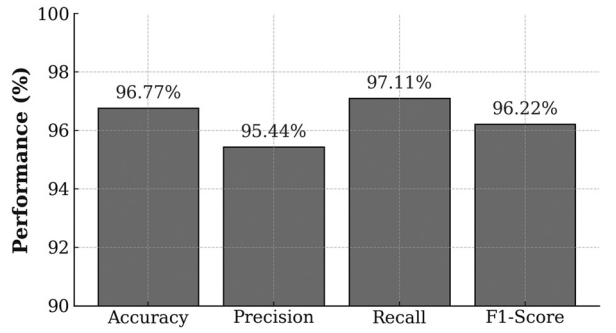

Fig. 13. GB model performance metrics for gear fault diagnosis.

behavior of acc3, ranging between –13.34 m/s^2 and 11.82 m/s^2, while acc4, shown in Fig. 10, shares the same value bounds as acc3 (–13.34 m/s^2 to 11.82 m/s^2), suggesting similar dynamics and possible sensor overlap in their placement. These figures collectively provide an overview of signal magnitude variation across the vibration sensors under different gear health states.

Figure 10 depicts the statistical behavior of acc2 across the ten classes. In Fig. 11, the SD of acc2 over 1000-sample intervals reveals the variability in signal strength, with values ranging from a minimum of 2.58 to a maximum of 4.30, clearly distinguishing gear states with stable versus fluctuating vibration profiles. Figure 12 shows the S of acc2, representing signal asymmetry, with values between –0.12 and 0.10. These statistical measures enhance the interpretability of vibration signals and are useful to identify fault characteristics that are not obvious in the time domain alone.

Figure 13 presents the evaluation results of the GB model applied to the processed Aalto Gear Fault Dataset. The GB classifier demonstrated strong diagnostic capability with an accuracy of 96.77%, indicating the overall correctness of predictions across all gear condition classes. The precision reached 95.44%, reflecting the model's ability to minimize false positives, which is critical in industrial fault scenarios where misdiagnosis could lead to unnecessary maintenance. The recall, at 97.11%, highlights the model's effectiveness in correctly identifying actual fault conditions, minimizing the risk of overlooking potential failures. The F1-score of 96.22%, as the harmonic mean of precision and recall, confirms a balanced performance.

As shown in Table V, the proposed GB model achieved the highest accuracy at 96.77 percent, followed by Artificial Neural Network at 93.54 percent and Random Forest at 92.78 percent. Support Vector Machine and *k*-Nearest Neighbors reached 90.15 and 88.32 percent accuracy, respectively, while logistic regression performed the lowest at 85.67 percent. This comparison confirms the superior performance of GB in extracting patterns from low-dimensional statistical features.

Figure 14 illustrates the confusion matrix summarizing the classification outcomes of the GB model across 10 gear condition classes, each with 30 test samples. The matrix reveals highly accurate classification results with dominant values along the diagonal. The Failure1 class achieved 29 correct predictions, with 1 misclassified as Failure2. Failure2 was predicted perfectly with 30 out of 30 correct, while Failure3 saw 28 correctly classified, with 1 misclassified as Failure4 and 1 as Failure6. Failure4 achieved 29 correct predictions, with 1 misclassified as Failure3. Failure5 also reached 30 correct predictions. For Failure6, 28 predictions were correct, with 2 misclassified as Failure8. Failure7 resulted in 29 correct predictions, with 1 instance misclassified as Failure9. Failure8 had 28 correct classifications, with 2 incorrectly identified as Failure6. Failure9 reached 29 accurate predictions, with 1 misclassified as Failure7. Finally, the healthy class achieved 30 out of 30 correct

Table V. Performance comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	85.67	84.12	86.43	85.26
Support Vector Machine	90.15	89.45	91.02	90.23
Random Forest	92.78	91.67	93.11	92.38
k-Nearest Neighbors	88.32	87.10	89.05	88.06
Artificial Neural Network	93.54	92.88	94.01	93.44
GB	**96.77**	**95.44**	**97.11**	**96.22**

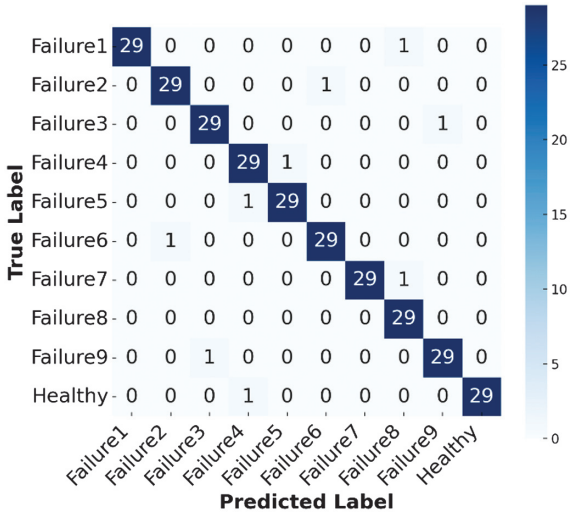

Fig. 14. Confusion matrix of the GB classifier for 10-class gear fault diagnosis.

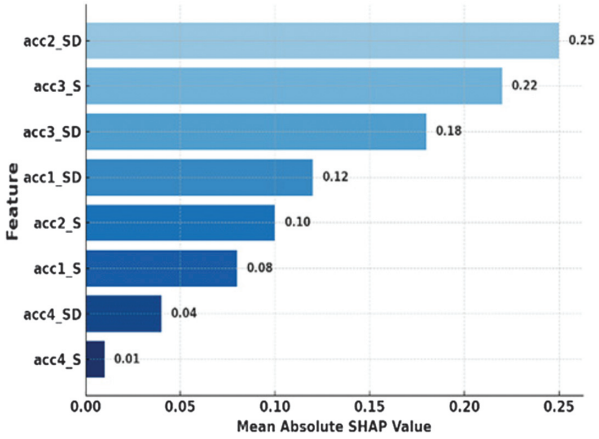

Fig. 15. SHAP summary plot for GB model.

classifications. The total number of correct predictions across all classes is 290 out of 300, which reflects a confusion matrix aligned with an overall classification accuracy of 96.77%. The misclassifications between Failure3 and Failure6, as well as between Failure6 and Failure8, likely stem from similarities in their vibration signatures, particularly in SD values, which can overlap when fault severity and frequency content are comparable. These overlaps suggest that while the GB model is highly accurate overall, subtle spectral and statistical similarities between certain fault types may require additional discriminative features to further reduce cross-class confusion.

Figure 15 presents the SHAP summary plot for the GB model, illustrating the mean absolute contribution of each input feature to the model's predictions across all gear fault classes. Among the eight features, acc2_SD emerged as the most influential, with a mean SHAP value of 0.25, followed closely by acc3_S at 0.22. These results indicate that asymmetry in the vibration signals from accelerometers 2 and 3 plays a key role in distinguishing between different gear conditions. Additionally, acc3_SD (0.18) and acc2_S (0.10) also showed moderate influence, further validating the diagnostic sensitivity of those sensor locations. In contrast, features such as acc4_S and acc4_SD had minimal impact, with SHAP values of 0.01 and 0.04, respectively, which suggests their limited contribution to fault discrimination.

This interpretability analysis highlights which features are most critical and reinforces the physical relevance of sensor placement and statistical signal behavior in gear fault diagnosis. The interpretability of SHAP outputs stems from their alignment with physically meaningful features whose contributions to fault diagnosis are understood by domain experts due to their established relevance in vibration analysis.

While the proposed XAI framework demonstrates high accuracy and interpretability on the ASD subset, its reliance on synthetic faults limits direct generalization to field conditions; future validation on real AGFD faults and in situ data is required. Moreover, TreeSHAP computation may become costly for larger datasets or streaming scenarios. To mitigate these issues, deployment should consider (i) hybrid pipelines that combine GB with lightweight noise-robust preprocessing or anomaly gating, (ii) batched or approximate SHAP computation (e.g., sampling-based explanations) for throughput, and (iii) an expanded roadmap that incorporates multi-modal fusion (e.g., vibration + acoustics + torque) and online learning to adapt to distribution shifts.

V. CONCLUSION

This study introduced an intelligent and interpretable framework for gear fault diagnosis using GB applied to the Aalto Gear Fault Dataset. By focusing on vibration signals from four accelerometers and extracting interval-based statistical features—SD and skewness—the model achieved high diagnostic performance, with an accuracy of 96.77%, precision of 95.44%, recall of 97.11%, and an F1-score of 96.22%. To bridge the gap between predictive power and transparency, SHAP was integrated to provide explainable insights into feature contributions, forming a robust XAI pipeline.

While the current study focused solely on time-domain statistical features and a single classifier, future work could expand by incorporating frequency-domain features, more complex ensemble models, and real-time implementation in industrial settings. Additionally, extending the framework to multi-sensor fusion and online adaptive learning would further enhance its applicability and reliability. Additionally, future studies should focus on validating the practical utility of the XAI framework by the incorporation of domain expert evaluations, real-time deployment scenarios, and comparative analysis with traditional diagnostic methods to ensure actionable and trustworthy decision-making.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

[1] A. Kuzmin, A. Ashori, P. Pantyukhov, Y. Zhou, L. Guan, and C. Hu, "Mechanical, thermal, and water absorption properties of HDPE/barley straw composites incorporating waste rubber," *Sci. Rep.*, vol. 14, no. 1, p. 25232, 2024.

[2] A. Baazaoui, S. Msolli, J. Alexis, O. Dalverny, and H. S. Kim, "Exploring joining techniques for diamond chips on metallized substrates: micro- and nano-scale mechanical testing approach," *Next Mater.*, vol. 7, p. 100349, 2025.

[3] L. A. Al-Haddad et al., "Energy consumption and efficiency degradation predictive analysis in unmanned aerial vehicle batteries using deep neural networks," *Adv. Sci. Technol. Res. J.*, vol. 19, no. 5, pp. 21–30, 2025.

[4] L. A. Al-Haddad, W. Giernacki, A. A. Shandookh, A. A. Jaber, and R. Puchalski, "Vibration signal processing for multirotor UAVs fault diagnosis: filtering or multiresolution analysis?," *Eksploat. Niezawodn. – Maint. Reliab.*, vol. 25, no. 1, p. 176318, 2023.

[5] L. A. Al-Haddad, W. Giernacki, A. Basem, Z. H. Khan, A. A. Jaber, and S. A. Al-Haddad, "UAV propeller fault diagnosis using deep learning of non-traditional χ^2-selected Taguchi method-tested Lempel–Ziv complexity and Teager–Kaiser energy features," *Sci. Rep.*, vol. 14, no. 1, p. 18599, 2024.

[6] A. A. Jaber, "Diagnosis of bearing faults using temporal vibration signals: a comparative study of machine learning models with feature selection techniques," *J. Fail. Anal. Prev.*, vol. 24, pp. 752–768, 2024.

[7] A. A. Jaber and L. A. Al-Haddad, "Integration of discrete wavelet and fast Fourier transforms for quadcopter fault diagnosis," *Exp. Tech.*, vol. 48, pp. 865–876, 2024.

[8] M. Irfan et al., "Revolutionizing wind turbine fault diagnosis on supervisory control and data acquisition system with transparent artificial intelligence," *Int. J. Green Energy*, vol. 22, no. 10, pp. 1–17, 2024.

[9] X. Yang, A. Jiang, W. Jiang, Y. Zhao, E. Tang, and S. Chang, "Abnormal detection and fault diagnosis of adjustment hydraulic servomotor based on genetic algorithm to optimize support vector data description with negative samples and one-dimensional convolutional neural network," *Mach.*, vol. 12, no. 6, p. 368, 2024.

[10] B. G. Mejbel, S. A. Sarow, M. T. Al-Sharify, L. A. Al-Haddad, A. A. F. Ogaili, and Z. T. Al-Sharify, "A data fusion analysis and random forest learning for enhanced control and failure diagnosis in rotating machinery," *J. Fail. Anal. Prev.*, vol. 24, pp. 2979–2989, 2024.

[11] A. A. F. Ogaili, M. N. Hamzah, and A. A. Jaber, "Enhanced fault detection of wind turbine using eXtreme gradient boosting technique based on nonstationary vibration analysis," *J. Fail. Anal. Prev.*, vol. 24, no. 2, pp. 877–895, 2024.

[12] G. Singh, M. Pal, Y. Yadav, and T. Singla, "Deep neural network-based predictive modeling of road accidents," *Neural Comput. Appl.*, vol. 32, no. 16, pp. 12417–12426, 2020.

[13] X. Liu, D. He, G. Lodewijks, Y. Pang, and J. Mei, "Integrated decision making for predictive maintenance of belt conveyor systems," *Reliab. Eng. Syst. Saf.*, vol. 188, pp. 347–351, 2019.

[14] M. Y. Arafat, M. J. Hossain, and M. M. Alam, "Machine learning scopes on microgrid predictive maintenance: potential frameworks, challenges, and prospects," *Renew. Sustain. Energy Rev.*, vol. 190, p. 114088, 2024.

[15] L. A. Al-Haddad, A. A. Jaber, M. N. Hamzah, and M. A. Fayad, "Vibration-current data fusion and gradient boosting classifier for enhanced stator fault diagnosis in three-phase permanent magnet synchronous motors," *Electr. Eng.*, vol. 106, no. 3, pp. 3253–3268, 2023.

[16] W. Guo, G. Wang, C. Wang, and Y. Wang, "Distribution network topology identification based on gradient boosting decision tree and attribute weighted naive Bayes," *Energy Rep.*, vol. 9, pp. 727–736, 2023.

[17] J. Yoon, "Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach," *Comput. Econ.*, vol. 57, no. 1, pp. 247–265, 2021.

[18] W. H. Alawee, L. A. Al-Haddad, A. Basem, D. J. Jasim, H. S. Majdi, and A. J. Sultan, "Forecasting sustainable water production in convex tubular solar stills using gradient boosting analysis," *Desalin. Water Treat.*, vol. 318, p. 100344, 2024.

[19] C. Zhang, H. Chen, X. Xu, Y. Duan, and G. Wang, "A data-driven metric-based proper orthogonal decomposition method with shapley additive explanations for aerodynamic shape inverse design optimization," *Adv. Eng. Inform.*, vol. 65, p. 103277, 2025.

[20] S. Lin, D. Song, B. Cao, X. Gu, and J. Li, "Credit risk assessment of automobile loans using machine learning-based SHapley additive exPlanations approach," *Eng. Appl. Artif. Intell.*, vol. 147, p. 110236, 2025.

[21] K. Hamad et al., "Explainable artificial intelligence visions on incident duration using eXtreme gradient boosting and SHapley additive exPlanations," *Multimodal Transp.*, vol. 4, no. 2, p. 100209, 2025.

[22] Z. Dahl et al., "Aalto gear fault datasets for deep-learning based diagnosis," *Data Brief*, vol. 57, p. 111171, 2024.

[23] W. H. Alawee, A. Basem, and L. A. Al-Haddad, "Advancing biomedical engineering: leveraging Hjorth features for electroencephalography signal analysis," *J. Electr. Bioimpedance*, vol. 14, no. 1, pp. 66–72, 2023.

[24] A. A. F. Ogaili, A. A. Jaber, and M. N. Hamzah, "Statistically optimal vibration feature selection for fault diagnosis in wind turbine blade," *Int. J. Renew. Energy Res. (IJRER)*, vol. 13, no. 3, pp. 1082–1092, 2023.

[25] J. Prawin, "Deep learning neural networks with input processing for vibration-based bearing fault diagnosis under imbalanced data conditions," *Struct. Health Monit.*, vol. 24, no. 2, pp. 883–908, 2024.

[26] L. Feng et al., "Scraper conveyor gearbox fault diagnosis based on multi-source heterogeneous data fusion," *Meas.*, vol. 247, p. 116797, 2025.

[27] G. Liu and L. Wu, "Running gear global composite fault diagnosis based on large model," *IEEE Trans. Ind. Inform.*, vol. 21, no. 5, pp. 4243–4251, 2025.

[28] M. R and R. R. Mutra, "Fault classification in rotor-bearing system using advanced signal processing and machine learning techniques," *Results Eng.*, vol. 25, p. 103892, 2025.

[29] K. N. Ravikumar, C. K. Madhusudana, H. Kumar, and K. V. Gangadharan, "Classification of gear faults in internal combustion (IC) engine gearbox using discrete wavelet transform features and K star algorithm," *Eng. Sci. Technol. Int. J.*, vol. 30, p. 101048, 2022.

[30] N. Zuber and R. Bajrić, "Gearbox faults feature selection and severity classification using machine learning," *Eksploat. Niezawodn. – Maint. Reliab.*, vol. 22, no. 4, pp. 748–756, 2020.

[31] J. Yu, S. Wang, L. Wang, and Y. Sun, "Gearbox fault diagnosis based on a fusion model of virtual physical model and data-driven method," *Mech. Syst. Signal Process.*, vol. 188, p. 109980, 2023.

[32] I. M. Jamadar et al., "Spur gear fault detection using design of experiments and support vector machine (SVM) algorithm," *J. Fail. Anal. Prev.*, vol. 23, no. 5, pp. 2014–2028, 2023.

[33] A. G. Abdulameer, A. S. Hammood, F. M. Abdulwahed, and A. A. Ayyash, "Naïve Bayes algorithm for timely fault diagnosis in helical gear transmissions using vibration signal analysis," *Int. J. Interact. Des. Manuf. (IJIDeM)*, vol. 19, no. 5, pp. 3695–3706, 2024.

[34] A. S. Hammood, A. G. Taki, N. S. Ibrahim, J. G. Mohammed, R. K. Jasim, and O. M. Jasim, "Optimizing failure diagnosis in helical gear transmissions with stochastic gradient descent logistic regression using vibration signal analysis for timely detection," *J. Fail. Anal. Prev.*, vol. 24, no. 1, pp. 71–82, 2023.

[35] M. Hu, S. Salmani Pour Avval, J. He, N. Yue, and R. M. Groves, "Explainable artificial intelligence study on bolt loosening detection using Lamb waves," *Mech. Syst. Signal Process.*, vol. 225, p. 112285, 2025.

[36] 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," *Mech. Syst. Signal Process.*, vol. 163, p. 108105, 2022.

[37] S. Gawde, S. Patil, S. Kumar, P. Kamat, K. Kotecha, and S. Alfarhood, "Explainable predictive maintenance of rotating machines using LIME, SHAP, PDP, ICE," *IEEE Access*, vol. 12, pp. 29345–29361, 2024.

[38] C. Men, Y. Han, P. Wang, J. Tao, and C.-G. Huang, "The interpretable reasoning and intelligent decision-making based on event knowledge graph with LLMs in fault diagnosis scenarios," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025.

[39] L. Lomazzi, S. Fabiano, M. Parziale, M. Giglio, and F. Cadini, "On the explainability of convolutional neural networks processing ultrasonic guided waves for damage diagnosis," *Mech. Syst. Signal Process.*, vol. 183, p. 109642, 2023.

[40] S. A. Khan, F. A. Khan, A. Jamil, and A. A. Hameed, "Interpretable motor sound classification for enhanced fault detection leveraging explainable AI," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, Mt. Pleasant, MI, USA, 2024, pp. 1–10.

[41] K. KN, A. Perrusquia, A. Tsourdos, and D. Ignatyev, "Integrating explainable AI into two-tier ML models for trustworthy aircraft landing gear fault diagnosis," in *AIAA SCITECH 2025 Forum*, Reston, VA, USA, 2025.