JOURNAL OF PRODUCTION ENGINEERING VOL. 27, NO. 2 (2024)
DOI: 10.24867/JPE-2024-02-001

Original research article

Journal of Production Engineering

https://jpe.ftn.uns.ac.rs/index.php/jpe
ISSN: 1821-4932 (print), ISSN: 2956-2252 (online)

Comparative analysis of gearbox fault detection using ensemble learning techniques with vibration sensor data

Nurudeen A. Raji 0000-0002-3967-9896, ***Rafiu O. Kuku*** 0000-0001-8537-5271
Abdullateef O. Bakare* 0009-0003-0197-1503, ***Medekannu M. Ogunbiyi*** 0009-0005-0829-1177
Tobiloba I. Morafa 0009-0009-1816-0583

Department of Mechanical Engineering, Lagos State University, Ojo (Main campus), Lagos State, Nigeria

ABSTRACT

Gearbox fault detection plays a crucial role in ensuring the reliable operation of machinery and preventing costly downtime. This research thesis aims to develop and evaluate ensemble learning techniques for accurate detection of gearbox broken tooth conditions using vibration data from SpectraQuest's Gearbox Fault Diagnostics Simulator. The dataset comprises vibration readings from sensors under both healthy and broken tooth conditions. A thorough analysis of the Gearbox Fault Diagnosis Dataset was conducted, integrating time and frequency domain analyses to inform feature engineering. A comprehensive comparative analysis of bagging, boosting, stacking, and voting approaches was conducted. The standout performer is the AdaBoostClassifierET, achieving an accuracy of 87.56%, precision of 88.36%, recall of 86.38%, and an F1 score of 87.36%. Bagging methods also exhibit commendable performance, with the BaggingClassifierET achieving an accuracy of 87.38%, precision of 87.17%, recall of 87.50%, and an F1 score of 87.34%. The research also highlights the significance of base model choices in ensemble techniques, as different base model choices yielded different results in all four techniques. The study surpasses previous work by incorporating a comprehensive set of ensemble techniques, advanced feature engineering informed by time and frequency domain analyses, and a nuanced evaluation of overfitting concerns.

ARTICLE INFO

Received: 2 April 2024
Revised: 30 May 2024
Accepted: 15 June 2024

KEYWORDS:
Vibration;
Fault diagnosis;
Gearbox;
Machine learning;
Detection;
Sensor.

**Corresponding author's.e-mail: bakare.opeyemi111@gmail.com*

1. INTRODUCTION

In mechanical engineering and industry, accurate and timely detection of gearbox faults is of paramount importance to ensure operational efficiency and avoid costly breakdowns. Over the years, researchers have endeavored to harness the power of advanced technologies and machine learning algorithms to develop more effective and efficient methods for gearbox fault diagnosis.

Gearbox fault detection plays a crucial role in ensuring the reliable operation of machinery and preventing costly downtime. In recent times, complex systems have become more interdependent, with various components such as bearings, gears, cam, and shafts relying on each other [1]. The failure of a single component can lead to the entire system shutting down, making gearbox failures a critical concern. Identifying and diagnosing faults in gearboxes is essential to mitigate potential monetary and life losses [2]. In this research thesis, we aim to investigate and propose machine learning techniques for gearbox fault detection to enhance system reliability and minimize downtime. Gearboxes operate under both constant and varying operating conditions. The continuous degradation of gears, particularly under varying operating conditions, poses a significant challenge. If gear faults go undetected, it can lead to severe consequences, including substantial financial losses and potential risks to human life [2].

In gear systems, stresses are primarily pure rolling at the pitch line, while rolling-sliding action occurs above and below the pitch line, with the sliding direction being opposite [3]. Adequate lubrication is crucial to ensure smooth operation in the sliding interfaces. However, insufficient lubrication can lead to direct contact between

Published by the University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia.

surfaces, resulting in surface disparities, increased temperature, and adhesive bonding under high pressure, ultimately leading to the breakdown of gear surfaces.
The gear roots experience both tension and compression simultaneously, with the root being the point of highest stress in tension [4]. The bending strength of the root depends on factors such as surface hardness, surface smoothness, sharpness of radius, and the presence of faults such as cracks or pitting [2]. Understanding these factors is essential for developing effective gearbox fault detection techniques [5]. Gearbox failures can be categorized into two main types: lubricated failures and non-lubricated failures [2]. Lubricated failures encompass issues such as pitting and mild wear, often caused by insufficient lubrication or adverse operating conditions [6]. On the other hand, non-lubricated failures include fractures and bending, resulting from excessive loads and harsh environmental conditions.
This study seeks to develop and evaluate ensemble learning techniques for the accurate and timely detection of gearbox broken tooth conditions using vibration data. Specifically, the study will utilize the Gearbox Fault Diagnosis Dataset recorded by SpectraQuest's Gearbox Fault Diagnostics Simulator. The study proposes a reliable and efficient machine learning ensemble-based approach that utilizes vibration data. The utilization of the Gearbox Fault Diagnosis Dataset recorded by SpectraQuest's Gearbox Fault Diagnostics Simulator will provide valuable insights into the effectiveness of the developed techniques.

2. METHODOLOGY

The process involves data collection, preprocessing, feature engineering, model selection, training, evaluation, and interpretation. The tool used to carry out every step of the experimentation and methodology was Python V3.10.1 and the development environment was a kaggle jupyter notebook Integrated Development Environment. The dataset employed in this research encompasses vibration data recorded by SpectraQuest's Gearbox Fault Diagnostics Simulator [7]. The dataset includes readings from four vibration sensors placed in different directions under varying load conditions. Two distinct scenarios were considered, the Healthy Condition represented by filenames starting with "h" (e.g., h30hz0.csv, h30hz10.csv, ..., h30hz90.csv). These files pertain to a healthy gearbox at different load levels ranging from 0% to 90% in increments of 10%. And the Broken Tooth Condition indicated by filenames starting with "b" (e.g., b30hz0.csv, b30hz10.csv, ..., b30hz90.csv). These files correspond to a gearbox with a broken tooth under varying load conditions.

2.1 Sensor configuration and data format

Data was collected from four sensors, denoted as a1, a2, a3, and a4, each providing readings in different directions. The data was sampled at a rate of 30Hz, as indicated by the "30hz" in the filenames. This information allows us to interpret the data as a continuous signal and analyze it both in the time domain and frequency domain.

The filename structure provides essential details of the prefix and suffix. The prefix is the first character in the filename denoting the gearbox condition, with "h" for healthy and "b" for a gearbox with a broken tooth and the suffix is the last 1-2 characters representing the load condition during data collection, ranging from 0 to 90 in increments of 10.
The dataset comprises a total of 8,036 samples, with an almost equal distribution between the two classes: healthy gearboxes (50.2%) and those with a broken tooth (49.8%). 80-20 splits of the dataset into training and testing sets were conducted. This stratified split ensures a proportional representation of both classes in both sets.

2.2 Data collection process

The data collection process involved reading and consolidating information from 20 files, ten each for healthy and broken tooth gearboxes. Each file corresponds to a specific load condition, contributing to a comprehensive dataset for analysis.
To facilitate analysis, the dataset was transformed into a "tall-form" where sensor readings are organized as follows: Sample Index (an index indicating the order of readings), State (indicating the gearbox condition i.e., healthy or broken tooth), Load (denoting the load condition during data collection), Sensor (specifying the sensor i.e., a1, a2, a3, or a4) from which the reading originated, and Reading (the actual vibration reading recorded by the respective sensor).
This melted dataset structure allows for efficient exploration and analysis of the sensor readings in subsequent stages of the research.

2.3 Time domain analysis

The data collected from sensor 'a1' was utilized to discern patterns and variations between healthy and broken tooth gearboxes under different load conditions. Readings from sensor 'a1' for both healthy and broken tooth gearboxes at loads of 0% and 90% were plotted.
To advance the analysis, features were extracted for each sample, encompassing sensor information, load, mean, standard deviation, kurtosis, skewness, and moments. The resulting dataset comprises a total of 8,036 samples, with the fault class constituting approximately 49.8% of the total samples. The dataset was then split into training and test sets.
The time domain analysis unveiled distinct patterns in sensor readings between healthy and broken tooth gearboxes. The subsequent feature generation process sets the stage for machine learning approaches to differentiate between the two gearbox states based on temporal characteristics.

2.4 Frequency domain analysis

In pursuit of a more comprehensive understanding of the gearbox fault diagnosis dataset, we turned to frequency domain analysis. This approach is crucial as it allows for uncovering valuable insights into the vibrational

characteristics of the system by examining its component frequencies.

Fourier transformation, conducted using the Scipy fast Fourier transform (FFT) module, was employed to decompose a single sensor reading into its constituent frequencies, providing a spectral representation of the data. From the resulting frequency spectrum, distinct peaks corresponding to specific frequencies were obtained. Furthermore, Power Spectral Density (PSD) analysis was employed to illustrate the distribution of power across different frequencies. This analysis provided a clearer representation of the dominant frequency components present in the data. Understanding these frequency components is essential for identifying anomalies or patterns indicative of gearbox faults, thereby contributing to more effective fault diagnosis and maintenance strategies.

2.5 Model selection

Different ensemble learning models were considered, including boosting, bagging, voting, and stacking algorithms. The selection of these models was based on their suitability for binary classification and their ability to handle high-dimensional feature sets. The chosen models were then trained on the training dataset. Feature matrices and target labels were fed into the models for parameter estimation, as illustrated in Fig. 1.

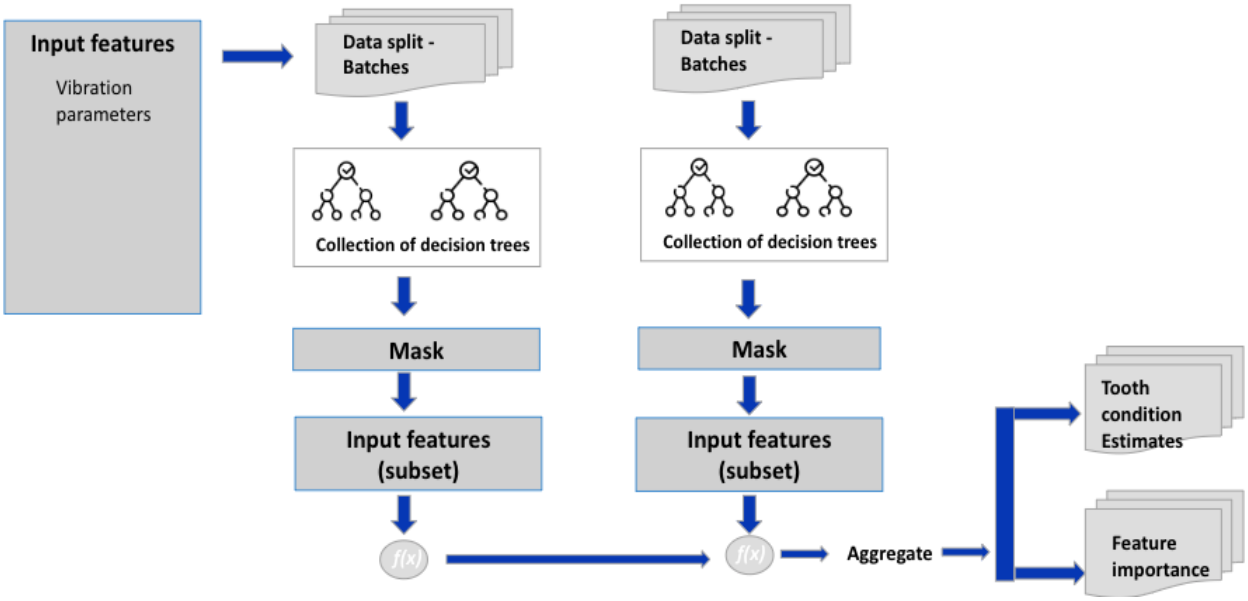

Fig. 1 *Model Training Workflow*

2.6 Model evaluation

The evaluation process aimed to provide a thorough understanding of each model's strengths and weaknesses, aiding in the selection of the most suitable approach for practical applications. Standard classification metrics used to evaluate the models include the accuracy metric, as expressed in Eq. (1). Accuracy measures the overall correctness of the model's predictions and is calculated as the ratio of correctly predicted instances to the total instances.

$$Accuracy = \frac{TruePositives+TrueNegatives}{TotalInstances} \quad (1)$$

The precision measures the accuracy of positive predictions and is defined as the ratio of true positive predictions to the total predicted positives as expressed in Eq. (2).

$$Precision = \frac{TruePositives}{TruePositives+FalsePositives} \quad (2)$$

Recall also known as sensitivity or true positive rate was used to quantify the ability of the model to capture all positive instances. It is calculated as the ratio of true positives to the total actual positives as expressed in Eq. (3).

$$Recall = \frac{TruePositives}{TruePositives+FalseNegatives} \quad (3)$$

The F1 score expressed in Eq. (4) is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

The confusion matrix offers a detailed breakdown of the model's predictions, including true positives, true negatives, false positives, and false negatives. It provides insights into the model's ability to classify instances correctly.

3. RESULTS AND DISCUSSION

This section presents a comprehensive analysis of the gearbox fault detection process, divided into three main parts: Time Domain Analysis, Frequency Domain Analysis, and Model Evaluation. Each part provides detailed insights into the methodologies and findings relevant to detecting gearbox faults using vibration sensor data. The subsequent comparative analysis summarizes the performance of various ensemble learning models.

3.1 Time domain analysis results

The analysis began by examining readings from sensor a1, which revealed noticeable differences in amplitude between the two gearbox states, indicating distinct vibration patterns (Fig. 2). To assess the overall distribution of readings across all sensors, boxplots were utilized (Fig. 3). Notably, sensor 'a1' exhibited a substantial amplitude difference between healthy and broken gearboxes, particularly under increasing load. In contrast, sensors 'a2', 'a3', and 'a4' showed comparatively smaller differences.

Focusing on sensor 'a1' in the first row of Fig. 3, a significant difference in amplitude distribution between the healthy and broken tooth gearboxes is observed. However, when excluding sensor 'a1' from the analysis, the standard deviation across all sensors aligned more closely (4.28) as opposed to 4.82 when sensor a1 is included. Hence, a decision was made to exclude sensor a1 from the analysis.

3.2 Frequency domain results

Upon visualizing the transformed data, we identified distinct peaks in the frequency spectrum (Red dots in Fig. 4). Primary harmonics emerged at 2.23 Hz intervals, accompanied by minor peaks between these intervals. This finding laid the groundwork for a more detailed analysis of spectral features.

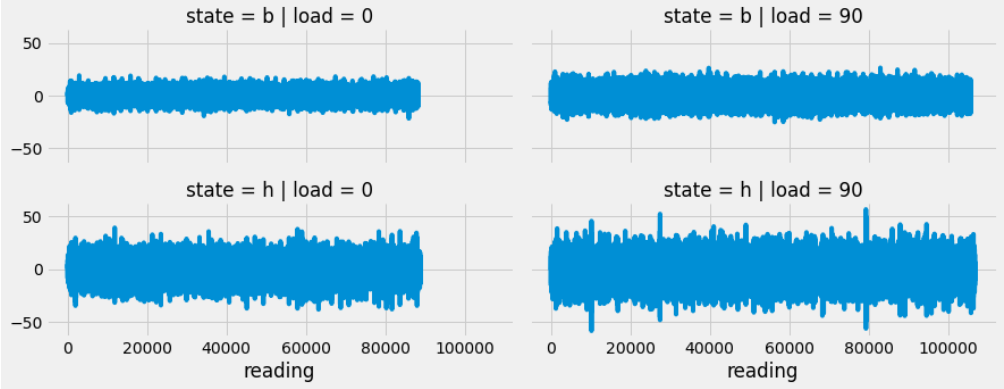

Fig. 2 *Readings from sensor a1*

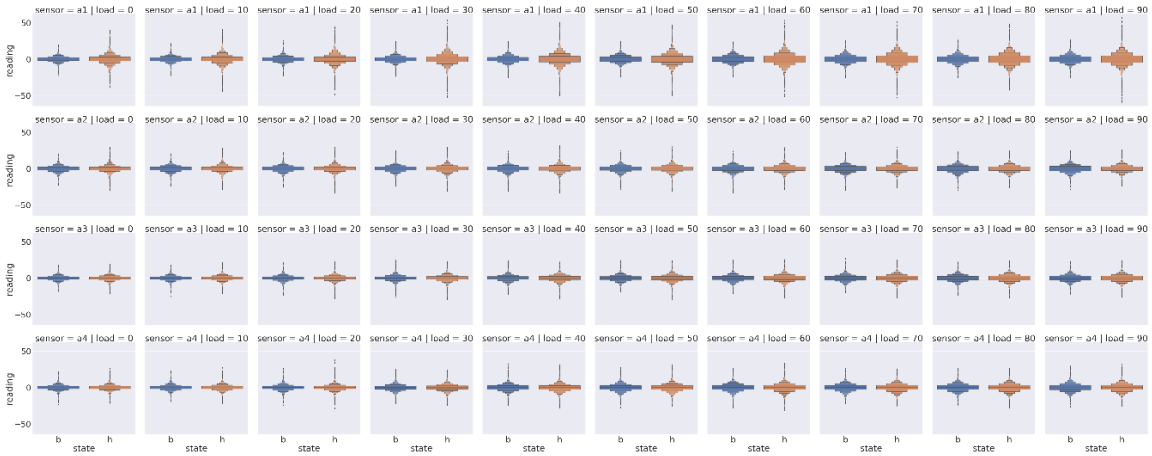

Fig. 3 *Plots of Sensor Readings for sensors a1, a2, a3, a4. Each row is a different sensor, each column shows increasing load and each plot shows the distribution of reading values between the healthy and broken gearbox*

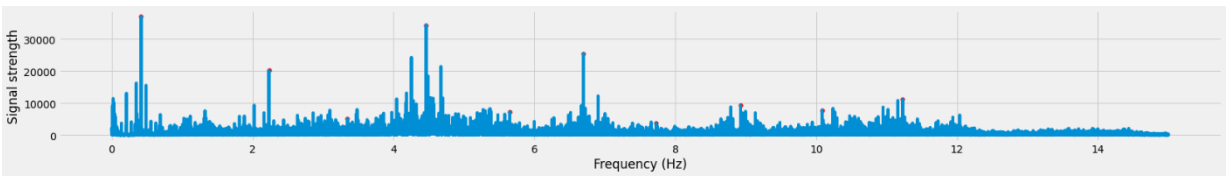

Fig. 4 *Peaks in the Transformed Data*

Expanding the investigation, the PSD for each sensor was compared at various loads between healthy and broken gearboxes. Discrepancies in power distribution, particularly on sensors a1, a2, and a3, were vividly displayed in the resulting plots, suggesting potential indicators of gearbox health. To assess temporal variations, spectrograph view could be employed to reveal the stability of the spectrum over time. The stability of the spectrum over time was demonstrated by the spectrogram plots, showing that the spectrum remains relatively constant across different time intervals.

Crucial information about the vibrational characteristics of the gearbox data is unveiled by Frequency domain analysis. Spectral features are explored, peaks identified, and PSDs compared, laying the groundwork for subsequent machine learning endeavors in gearbox fault detection. The robustness of these analyses will be enhanced by the stability of the spectrum over time, contributing valuable insights to the expected output.

3.3 Model evaluation

Table 1 presents the modeling results of boosting algorithms, offering a comprehensive overview of performance metrics for various ensemble learning techniques utilized in gearbox fault detection using vibration sensor data. The study specifically concentrates on Gradient Boosting Classifier, XGB Classifier, AdaBoost Classifier ET, AdaBoostClassifierRF, and AdaBoostClassifierBG. The results highlight the effectiveness of boosting ensemble learning techniques in enhancing gearbox fault detection model performance.

AdaBoostClassifierET, employing the extra trees algorithm as the base estimator, emerges as noteworthy, achieving the highest accuracy (87.56%) and precision (88.36%). This indicates its capability to accurately identify positive instances while minimizing false positives. AdaBoostClassifierBG, utilizing the Bagging Algorithm as the base estimator, also demonstrates balanced performance across all metrics, underscoring its reliability in fault diagnosis.

Table 1 - *Modeling results of Boosting Algorithms.*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Gradient Boosting Classifier	83.83	83.33	84.38	83.85
XGB Classifier	85.32	85.79	84.5	85.14
AdaBoostClassifierET	87.56	88.36	86.38	87.36
AdaBoostClassifierRF	86.75	87.39	85.75	86.56
AdaBoostClassifierBG	87.44	87.38	87.38	87.38

Fig. 5 illustrates the confusion matrix for the top-performing boosting algorithm, the AdaBoost Classifier ET, in the context of gearbox fault detection. In this context, the positive label indicates a fault condition, specifically a broken tooth in the gearbox, while the negative label denotes a healthy tooth condition. The classifier achieves an accuracy of 87.56%, indicating its proficiency in making correct predictions.

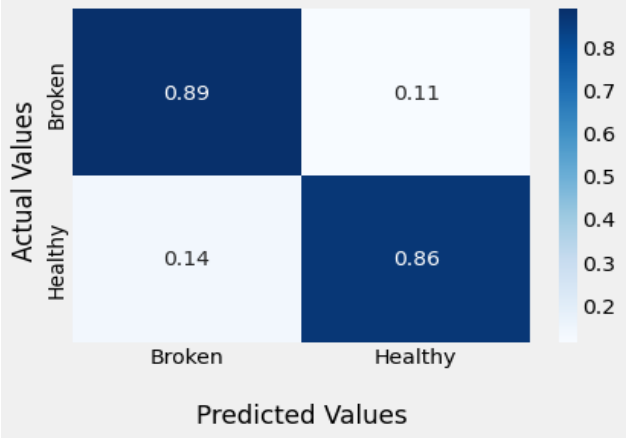

Fig. 5 *Confusion Matrix - AdaBoost Classifier ET*

Within the matrix, the true negatives (88.74%) represent instances where the model accurately identified healthy conditions, while the false positives (11.26%) indicate cases where the classifier incorrectly flagged a healthy condition as faulty. On the fault side, the false negatives (13.63%) depict instances where the model failed to detect an actual fault, and the true positives (86.38%) highlight successful identifications of faulty conditions. Analyzing these components underscores the importance of scrutinizing false positives and false negatives. Minimizing these errors is critical for refining precision (correctly identifying faults when predicted) and recall (capturing all actual faults), respectively.

This examination of the confusion matrix provides nuanced insights into the AdaBoost Classifier ET's performance, offering valuable information for optimizing the model to enhance gearbox fault detection, particularly in distinguishing between healthy and faulty tooth conditions in industrial machinery.

In Table 2, the performance metrics of the Bagging Classifier models, BaggingClassifierRF and BaggingClassifierET, are elaborated within the context of gearbox fault detection. For BaggingClassifierRF, the model demonstrates an accuracy of 86.69%, with precision, recall, and F1 score at 87.09%, 86.00%, and 86.54%, respectively. These metrics collectively illustrate the model's proficiency in correctly identifying both healthy and faulty tooth conditions. Similarly, BaggingClassifierET exhibits strong performance, achieving an accuracy of 87.38%, precision of 87.17%, recall of 87.50%, and an F1 score of 87.34%. This comprehensive evaluation highlights the robustness of both Bagging Classifier models in effectively discerning between healthy and faulty tooth conditions in gearbox fault detection tasks.

Comparing the performance metrics of the Bagging Classifier models (BaggingClassifierRF and

BaggingClassifierET) with the boosting models from Table 1 reveals interesting insights into their respective strengths in gearbox fault detection. The boosting models, represented by AdaBoost Classifier ET, AdaBoostClassifierRF, and AdaBoostClassifierBG, consistently showcase competitive metrics. It could be observed that for instance, AdaBoost Classifier ET achieves an accuracy of 87.56%, slightly outperforming BaggingClassifierET. Additionally, in terms of precision, recall, and F1 score, the boosting models exhibit comparable or marginally superior performance compared to the Bagging Classifier models. This suggests that both ensemble techniques demonstrate efficacy in handling the complexities of gearbox fault diagnosis, with boosting models holding a slight edge in certain metrics.

Table 2 - *Performance Metrics of Bagging Classifier Models*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Bagging Classifier RF	86.69	87.09	86	86.54
Bagging Classifier ET	87.38	87.17	87.5	87.34

Table 3 presents the modeling results for the StackingClassifier, revealing its performance in gearbox fault detection using ensemble learning. The StackingClassifier achieves an accuracy of 85.88%, showcasing its ability to make correct overall predictions. The precision of 85.95% indicates its competence in accurately identifying positive instances, representing the cases of gearbox faults. The recall, at 85.62%, signifies the model's capability to capture the majority of actual positive instances, demonstrating its sensitivity to identifying faults. The F1 score, which combines precision and recall into a single metric, stands at 85.79%, underscoring the balanced performance of the StackingClassifier. These results collectively suggest that the StackingClassifier is a promising ensemble learning technique for gearbox fault detection, providing a harmonious trade-off between precision and recall, crucial metrics in the context of condition monitoring and fault diagnosis in mechanical systems.

Table 3 - *Performance Metrics of The Stacking Classifier*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
StackingClassifier	85.88	85.95	85.62	85.79

Table 4 presents the performance metrics of two different configurations of the VotingClassifier:

- VotingClassifierEBA (with ExtraTrees estimator
- Bagging estimator,
- AdaBoostModelET as base estimators) and VotingClassifierDKS (with Decision Tree Classifier,
- KNeighborsClassifier
- Support Vector Classifier as base estimators).

These models were evaluated for gearbox fault detection based on accuracy, precision, recall, and F1 score. The VotingClassifierEBA achieved an accuracy of 87.50%, with a precision of 87.86%, recall of 86.88%, and an F1 score of 87.37%. In comparison, the VotingClassifierDKS exhibited an accuracy of 85.88%, precision of 86.68%, recall of 84.62%, and an F1 score of 85.64%.

These results provide valuable insights into the comparative performance of the two configurations and show that the VotingClassifierEBA consistently outperforms the VotingClassifierDKS on all evaluated metrics. The high precision and recognition scores achieved by both configurations demonstrate their effectiveness in accurately identifying gear faults while minimizing false positives and false negatives. Furthermore, the robust performance of the VotingClassifierEBA highlights its superior ability to process the dataset and make reliable predictions. This suggests that the VotingClassifierEBA is the better choice for real-world applications where accuracy and reliability are critical.

Table 4 - *Performance Metrics of The Voting Classifier*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
VotingClassifierEBA	87.5	87.86	86.88	87.37
VotingClassifierDKS	85.88	86.68	84.62	85.64

The confusion matrix in Fig. 6 illustrates the performance of the best voting model, VotingClassifierEBA, for gearbox fault detection. In this matrix, the diagonal elements indicate correctly classified instances, with an accuracy of 87.50%. The model demonstrates a balanced ability to identify both positive and negative cases, as evident from the precision of 87.86% and recall of 86.88%. These metrics signify the model's effectiveness in correctly recognizing instances of gearbox faults while minimizing false positives and false negatives. The F1 score, calculated as the harmonic mean of precision and recall, stands at 87.37%, further emphasizing the model's overall performance. These results collectively highlight the robustness of the VotingClassifierEBA in gearbox fault diagnosis, showcasing its potential for practical applications in fault detection systems.

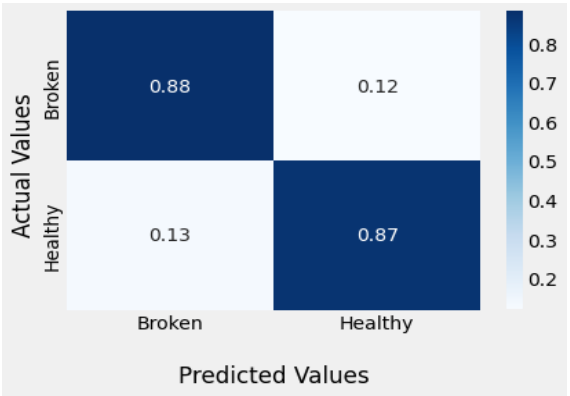

Fig. 6 *Confusion Matrix – VotingClassifierEBA*

3.4 Comparative analysis

In Fig. 7, the bar chart vividly illustrates the varying accuracies of different ensemble learning models employed for gearbox fault detection. The best models in each technique have been selected for this comparative analysis. Notably, the AdaBoostClassifierET emerges as the highest-performing model with an accuracy of 87.56%, showcasing its capability to make correct predictions across fault and non-fault instances. Following closely is the VotingClassifier at 87.50%, indicating its robust performance in identifying gearbox faults. The BaggingClassifierET and StackingClassifier exhibit competitive accuracies of 87.38% and 85.88%, respectively. This demonstrates the effectiveness of ensemble learning techniques in enhancing overall model accuracy, which is crucial for real-world applications in fault diagnosis.

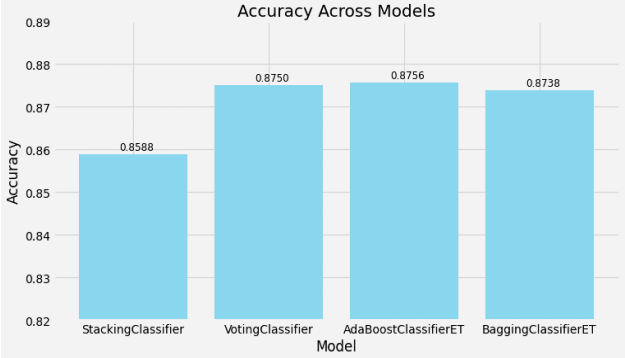

Fig. 7 *Bar Chart of Model Accuracies*

Turning attention to Fig. 8, the bar chart illustrates the recall values, providing insights into the models' abilities to correctly identify instances of gearbox faults. Here, the BaggingClassifierET stands out with a recall of 87.50%, underscoring its proficiency in capturing a significant proportion of actual positive cases. The VotingClassifier closely follows with a recall of 86.88%, indicating its balanced sensitivity to both fault and non-fault conditions. The AdaBoostClassifierET and StackingClassifier exhibit commendable recall values of 86.38% and 85.62%, respectively, emphasizing their effectiveness in minimizing false negatives and ensuring comprehensive fault detection.

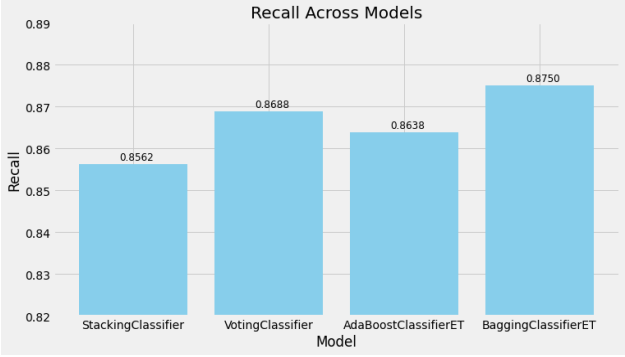

Fig. 8 *Bar Chart of Model Recalls*

Precision, visualized in Fig. 9, is a critical metric to assess the models' accuracy in identifying true positive cases among all predicted positives. The AdaBoostClassifierET impressively leads with a precision of 88.36%, showcasing its ability to minimize false positives. The VotingClassifier follows closely at 87.86%, emphasizing its precision in correctly classifying gearbox faults. The BaggingClassifierET and StackingClassifier exhibit competitive precision values of 87.17% and 85.95%, respectively. These results highlight the models' precision-driven performance, crucial for applications where minimizing false alarms is paramount.

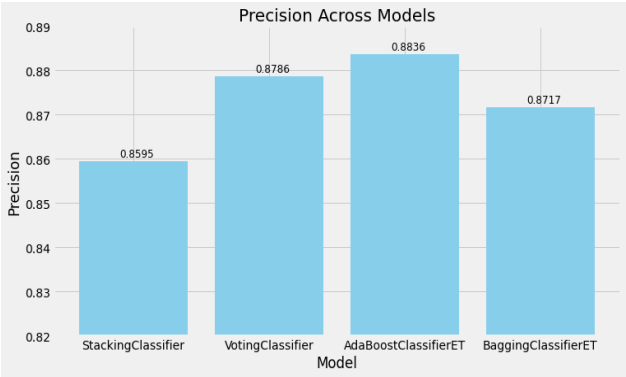

Fig. 9 *Bar Chart of Model Precisions*

Finally, Fig. 10 showcases the F1 scores, providing a holistic assessment by balancing precision and recall. The AdaBoostClassifierET again leads with an F1 score of 87.36%, emphasizing its comprehensive and balanced performance. The VotingClassifier closely follows at 87.37%, underlining its effectiveness in achieving a harmonious trade-off between precision and recall.

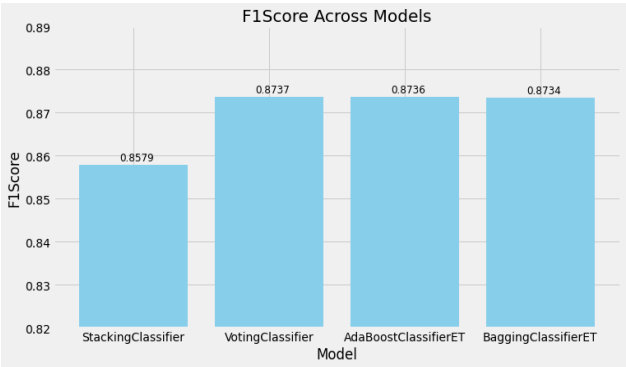

Fig. 10 *Bar Chart of Model F1-Scores*

The BaggingClassifierET and StackingClassifier present competitive F1 scores of 87.34% and 85.79%, respectively. Collectively, these findings substantiate the efficacy of ensemble learning models in gearbox fault detection, with each model demonstrating unique strengths in balancing precision, recall, and overall accuracy as shown in Table 5. Fig.11 presents the Receiver Operating Characteristic (ROC) curve for the best-performing model, AdaBoostClassifierET, depicting its discrimination ability

between healthy and faulty gearbox conditions. The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate at various thresholds. The Area Under the Curve (AUC) summarizes the ROC curve, indicating the model's ability to distinguish between classes.

Table 5 - *Summary Evaluation of Ensemble Models for Gearbox Fault Detection: Accuracies, Precisions, Recalls, and F1 Scores*

Model	Accuracy	Precision	Recall	F1 Score
	(%)	(%)	(%)	(%)
Stacking Classifier	85.88	85.95	85.62	85.79
AdaBoost Classifier ET	87.56	88.36	86.38	87.36
Bagging ClassifierET	87.38	87.17	87.5	87.34
Voting ClassifierEBA	87.5	87.86	86.88	87.37

The evaluation process was standardized across all models, ensuring consistent and fair comparisons. After training each model on the standardized training dataset, predictions were made on the test dataset. Subsequently, the models were subjected to the evaluation code.

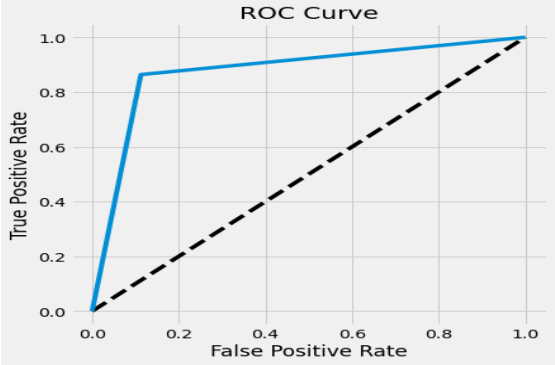

Fig. 11 *Receiver Operating Characteristic (ROC) curve for the best-performing model*

The evaluation process was standardized across all models, ensuring consistent and fair comparisons. After training each model on the standardized training dataset, predictions were made on the test dataset. Subsequently, the models were subjected to the evaluation code.

The ROC-AUC curve, along with the AUC score, adds a layer of analysis by illustrating the trade-off between true positive rate and false positive rate. This curve is particularly valuable for assessing the model's discriminative ability. Together, these evaluation metrics contribute to a holistic assessment, enabling informed decision-making regarding the most effective machine learning approach for gearbox fault diagnosis in industrial applications.

The curve gracefully ascends, with an Area Under the Curve (AUC) score of 0.8756, indicating strong predictive performance. The False Positive Rate (FPR) and True Positive Rate (TPR) trade-off is visually evident, showcasing the model's capability to balance between correctly identifying positive cases (faulty gearbox) and minimizing false alarms. The thresholds at various points on the curve are marked, offering insights into the model's sensitivity at different decision boundaries. The steep rise in TPR with a relatively low FPR underscores the model's effectiveness in distinguishing between classes. Overall, the ROC curve and AUC score affirm the robustness and reliability of the AdaBoostClassifierET in gearbox fault detection.

4. CONCLUSION

The study evaluates ensemble learning techniques for gearbox fault detection, primarily using vibration sensor data augmented by time and frequency domain analysis in a careful feature engineering process. The AdaBoostClassifierET performed best with 87.56% accuracy, 88.36% precision, 86.38% recall and 87.36% F1-score, making it the most reliable choice for distinguishing healthy from faulty gearboxes.

Other ensemble methods, especially bagging-based methods, also performed well. For example, the BaggingClassifierET achieved an accuracy of 87.38%, a precision of 87.17%, a recall of 87.50% and an F1-score of 87.34. The study emphasizes the importance of selecting appropriate base models in ensemble techniques and finds that stacking and voting methods perform excellently in combination with base models from boosting or bagging algorithms, providing valuable insights for practitioners in the detection of gearbox faults.

In contrast to previous work that achieved 87.5% accuracy with the logit boosting algorithm [8], our approach outperforms this value by exploring a wider range of ensemble techniques (bagging, boosting, stacking, and voting), allowing for a more comprehensive analysis. Notably, [8] reported a test accuracy of 100, possibly indicating overfitting and limited generalizability, highlighting the need for robust evaluation metrics.

Furthermore, the study is limited to broken teeth in gearboxes, which may limit direct generalization to other failure modes or operating contexts. Future research could explore adaptive ensemble configurations that are able to dynamically adjust the ensemble composition during runtime to account for evolving data patterns. This adaptive approach promises to improve fault detection and diagnosis in different operating scenarios and provide a more robust framework for monitoring and maintenance of transmissions.

REFERENCES

[1] Davis, J.R. (2005). Gear materials, properties, and manufacture. DOI: 10.31399/asm.tb.gmpm.9781627083454.

[2] Liang, X., Zuo, M.J., Feng, Z. (2018). Dynamic modeling of gearbox faults: A review. *Mechanical Systems and Signal Processing*, vol. 98, p. 852-876, DOI: 10.1016/j.ymssp.2017.05.024.

[3] Amarnath, M., Lee, S.-K. (2015). Assessment of surface contact fatigue failure in a spur geared system based on the tribological and vibration parameter analysis. *Measurement*, vol. 76, p. 32-44, DOI: 10.1016/j.measurement.2015.08.020.
[4] Mohammed, O.D., Rantatalo, M., Aidanpää, J.-O. (2015). Dynamic modelling of a one-stage spur gear system and vibration-based tooth crack detection analysis. *Mechanical Systems and Signal Processing*, vol. 54, p. 293-305, DOI: 10.1016/j.ymssp.2014.09.001.
[5] Bojanić Šejat, M., Rackov, M., Knežević, I., Živković, A. (2022). Modal analysis of ball bearings using finite element method. *Journal of Production Engineering*, vol. 25, no. 2, p. 20-24, DOI: 10.24867/JPE-2022-02-020.
[6] Sándor, B. (2022). Finite element method analysis of various tooth roots on the pinion of connecting helical gear pairs having complex teeth by constant moment. *Journal of Production Engineering*, vol. 25, no. 2, p. 13-19, DOI: 10.24867/JPE-2022-02-013.
[7] Pandya, Y. Gearbox fault diagnosis data, from *https://data.openei.org/submissions/623*, accessed on [Accessed: 24th July, 2023].
[8] Kankar, H., Prakash, J. (2023). Comparative analysis of ensemble learners for broken tooth diagnostics in gears. *Life Cycle Reliability and Safety Engineering*, vol. 12, no. 4, p. 277-284, DOI: 10.1007/s41872-023-00235-5.