An Ensemble Machine Learning Model for Predictive Maintenance on Water Injection Pumps in the Oil and Gas Industry

Salama Mohamed Almazrouei[1]*, Fikri Dweiri[2], Ridvan Aydin[3], Abdalla Alnaqbi[4]
Department of Industrial Engineering and Engineering Management-College of Engineering,
University of Sharjah, Sharjah, United Arab Emirates[1, 2, 3]
ADNOC Offshore, Abu Dhabi, United Arab Emirates[4]

***Abstract*—The effective operation of water injection pumps is vital for enhancing oil recovery in the oil and gas industry. To ensure optimal pump performance and prevent unplanned downtime, this study focused on implementing predictive maintenance strategies. We began by identifying five critical operational parameters—Seal Pressure 1, Seal Pressure 2, Vibration Data for the Drive End (VIB DE), Vibration Data for the Non-Drive End (VIB NDE), and Ampere. These parameters were monitored and analyzed to evaluate their impact on pump performance and maintenance needs. To achieve this, we applied three machine learning algorithms: Extreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LGBM), and Random Forest. Each algorithm was independently trained and tested on the dataset corresponding to each operational parameter. We assessed their performance using key accuracy metrics, including R squared, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Following this, we developed an Ensemble model, combining the predictive outputs of XGBoost, LGBM, and Random Forest. The Ensemble model was then applied to the same parameters to evaluate its ability to address the limitations observed in standalone models. The results demonstrated that the Ensemble model consistently delivered superior performance, achieving lower RMSE and MAE values and higher R squared coefficients across all parameters. This study culminates in the validation of the Ensemble model as a robust and reliable approach for predictive maintenance. By leveraging the strengths of multiple algorithms, the Ensemble model offers significant improvements in accuracy and reliability, contributing to more effective maintenance systems for the oil and gas industry.**

Keywords—Ensemble machine learning models; oil and gas industry; predictive maintenance; water injection pumps

I. INTRODUCTION

The oil and gas sector confronts the critical challenge of substantially reducing operational costs while upholding safety standards [1]. Fortuitously, artificial intelligence (AI) has become instrumental in addressing the challenges faced by the oil and gas industry (OGI), capitalizing on technological advancements and the Big Data revolution to facilitate informed decision making and expedite the transition from issue identification to execution [2]. Encompassing a range of operations spanning exploration to distribution, the OGI functions within a multifaceted environment characterized by extensive infrastructure and high-value assets [3]. Efficient maintenance and reliability management are imperative for ensuring optimal production, safety, and cost-effectiveness in this industry [3]. The adoption of predictive maintenance (PdM) emerges as a pivotal methodology, seamlessly integrating data analysis, machine learning (ML) algorithms, and sensor technologies to collect real-time operational and sensor data [4]. This proactive approach plays a crucial role in early failure identification, thereby mitigating the consequences of unforeseen downtime. A thorough examination of the existing literature reveals numerous successful AI implementations across various domains of petroleum engineering [5]. Within the OGI, a particular emphasis on PdM is evident, particularly concerning Water Injection Pumps (WIPs), which play a fundamental role in maintaining reservoir pressure and optimizing oil recovery [3]. By harnessing advanced technologies like ML and deep learning (DL), PdM for WIPs introduces enhanced strategies for predicting failures early and facilitating real-time condition-based proactive maintenance [6]. In contrast, traditional maintenance approaches often lead to unnecessary actions and disruptive costs, while reactive maintenance poses risks to safety and the environment [7]. PdM tackles these challenges by leveraging cutting-edge technologies and data analytics to continuously monitor the real-time health and performance of equipment. The comprehensive data-driven approach employed by PdM models, drawing on real-time data from sensors, control systems, and historical maintenance records, enables the early detection of potential issues [3,8]. This proactive intervention empowers organizations to anticipate and address impending challenges, effectively minimizing downtime, optimizing maintenance strategies, and improving operational efficiency [9]. The integration of PdM models establishes a proactive maintenance paradigm that fortifies reliability, reduces costs, and prolongs the lifecycle of critical assets [3,4]. The early identification and resolution of potential issues provide operators with the capability to circumvent costly breakdowns, mitigate production losses, and ensure uninterrupted operations. PdM also presents the opportunity for more efficient resource planning and allocation, enabling operators to optimize spare parts inventories and streamline maintenance schedules [2,4,5,10].

Digitalization and the Internet of Things (IoT) generate extensive data, driving the significance of Predictive Maintenance (PdM) in optimizing upstream rotating equipment in the Oil, Gas, and Petrochemical (OGP) industry [11, 12].

Efficient operation of Water Injection Pumps (WIPs) enhances oil recovery and operational success [1–3]. PdM minimizes unplanned downtime and improves pump performance, with Deep Learning (DL) playing a pivotal role in recent studies. Janssens et al. [13] used CNNs for health monitoring via infrared thermal images, showcasing potential in anomaly detection, while Sampaio et al. [14] employed ANNs to predict motor failures, albeit with limited performance analysis. Bekar et al. [15] utilized K-means and PCA for motor PdM, facing challenges related to motor type specificity. Falamarzi et al. [16] applied ANN and SVR to tram track gauge prediction, and Susto et al. [17] proposed a PdM system for epitaxy processes, lacking equipment-specific context.

In developing countries, implementing PdM in the OGP sector encounters barriers such as limited skilled personnel, sensor access, infrastructure constraints, financial limitations, and cultural challenges [18-19]. Initiatives for sustainability and diversification, alongside training and infrastructure investments, aim to address these challenges [19-20]. Despite advancements, there is a research gap in applying AI to predict failures in WIPs, critical for health, safety, and environmental outcomes. This study addresses the gap by leveraging ML models like XGBoost and LGBM to predict WIP failures, focusing on asset loss, regulatory compliance, corporate reputation, and production impacts [3–5].

These algorithms, recognized for their high performance, have not been extensively studied in conjunction with ensemble methods such as Random Forest. The Ensemble model demonstrates unparalleled accuracy and increases the accuracy of predictions. This research is driven by the urgency to provide advanced solutions for PdM in the OGI sector, addressing the complexities of diverse operational parameters. This study specifically addresses the maintenance of water injection pumps, focusing on critical factors such as currents, pressure, and vibrations because these factors are crucial for maintenance but are often under-represented in existing research. This study innovatively integrates multiple algorithms within ensemble models to enhance predictive accuracy and address maintenance challenges in water injection pumps. Unlike previous approaches that often lack specificity in algorithm selection for factors such as currents, pressure, and vibrations, a systematic approach is employed to optimize performance in real-world operational environments. This refinement distinguishes the work by effectively applying ensemble techniques to improve Prognostics and Health Management (PHM) systems, particularly in critical applications such as WIPs. The findings of this study are poised to significantly contribute to the field by presenting a holistic and enhanced approach to predictive modeling in industrial settings. In the upcoming sections, this study delves into PdM through ML. Section II explains the methodology, highlighting the significance of the Ensemble model. Section III presents the results and analysis, while discussion is given in Section IV. Limitation and future work is given in Section V and Section VI respectively. Finally, the paper is concluded in Section VII.

II. Methodology

Predictive maintenance (PdM) in the OGI industry is critical for optimal production, safety, and cost-effectiveness. This study employs an ensemble of ML models—XGBoost, LGBM, and Random Forest—to enhance predictive accuracy. The unique strengths of each algorithm contribute to a robust framework. The ensemble methodology leverages complementary features, addressing individual weaknesses and mitigating biases. This novel approach aims to outperform standalone models, offering more accurate and reliable results. Fig. 1 illustrates the methodology implemented in our study. The flowchart visually outlines the sequential steps involved: data collection from various sources relevant to WIPs performance, including currents, pressure, and vibrations; preprocessing of the collected data to handle missing values, normalize them, and prepare them for analysis; selection of pertinent features influencing WIPs performance; training of machine learning models, comprising individual algorithms and Ensemble models, using the preprocessed data; evaluation of model performance using metrics such as RMSE to determine accuracy; creation of an Ensemble model by combining outputs from multiple models to enhance predictive accuracy; validation of the Ensemble model using test data to ensure reliability; and deployment of the final model for real-time predictive maintenance of water injection pumps, enabling proactive maintenance scheduling. To model and simulate the water injection network system, the relationships between different components and their respective parameters are established using the given formulae and principles of fluid mechanics. This includes deriving equations for the pump model, the water distributing station model, the well model, the tube element model, the node element model, and the system model, as outlined below. The pump model is represented by the quadratic equation:

$$H = AQ^2 + BQ + C \tag{1}$$

where H is the pump outlet pressure, Q is the pump flow rate, and A, B, and C are co-efficient representing the quadratic, linear, and constant terms, respectively. The water distributing station and well model can be expressed as:

$$P = A\ Q + B \tag{2}$$

where P denotes the pressure at the water distributing station or injection well, Q is the flow rate, and A and B are the linear and constant term coefficients, respectively [21]. The tube element model is described by:

$$qi = ki(hk - hj) \tag{3}$$

$$hf\ = Pi - Pj \tag{4}$$

where qi is the flow rate through pipeline element i, ki is a variable related to the pipe-line element length, inner diameter, and friction coefficient, hk and hj are the gross heads at the two ends of the pipeline element, and hf is the pressure loss across the pipe sec-tion. The pressure loss in the pipe section is calculated using the Darcy formula:

$$hf = Lv^2/2dg \tag{5}$$

where *hf* is the pressure loss, *L* is the pipe section length, d is the diameter, *v* is the fluid flow velocity, and *g* is the gravitational acceleration. Using these relationships, the system of equations is formulated to describe the behavior of the water injection network. Numerical methods are employed to solve these equations iteratively, allowing for the simulation of network parameters such as pressure and flow rate at all nodes.

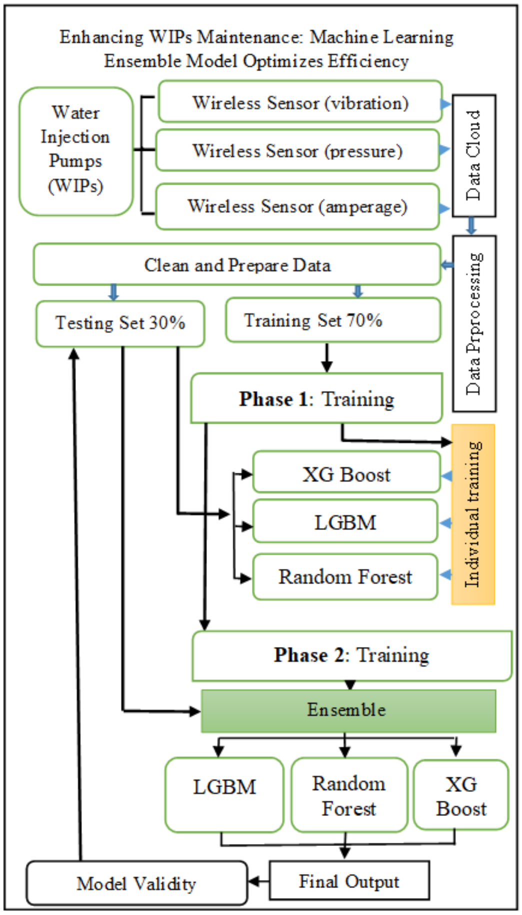

Fig. 1. Methodology flowchart.

A. *Ensemble Model Construction*

The present study incorporates a diverse ensemble of ML models; namely, X Boost, LGBM, and Random Forest. Each algorithm brings unique strengths, contributing to the overall robustness of the predictive framework. XGBoost's scalability, LGBM's high-performance gradient boosting, and Random Forest's ensemble learning approach individually address distinct aspects of the predictive task. The ensemble methodology aims to leverage the complementary strengths of each algorithm, enhancing predictive accuracy. By fusing the predictive capabilities of XGBoost, LGBM, and Random Forest, the study seeks to capitalize on their collective intelligence, mitigating individual weaknesses. The study asserts that this amalgamation, orchestrated through ensemble techniques, can yield more accurate and reliable results compared to each algorithm operating in isolation. This approach is grounded in the empirical observation that ensemble models often outperform individual algorithms by mitigating biases and reducing overfitting.

B. *Data Collection and Preprocessing*

The data utilized in this study originates from sensor readings, maintenance records, and operational parameters, collectively offering insights into the pump system's behavior. Sensor data provide real-time insights into operational aspects, including pressure levels, temperatures, and vibrations. Maintenance records offer a historical perspective, detailing interventions over time. Table I presents a succinct overview of the key operational parameters meticulously chosen for the analysis of the Work in Progress WIPs system. Among these parameters are the Ampere, denoting the current employed by the pump; VibDE, representing the vibration in the Drive End bearing; VibNDE denoting the vibration in the Non-Drive End bearing; and 1st Press, and 2nd Press delineating the 1st and 2nd Stage Seal Pressures, respectively.

TABLE I. LIST OF SELECTED RUNNING PARAMETERS FOR WIPS SYSTEM ANALYSIS

Name	Description
1st_Press	1st Stage Seal Press
2nd_Press	2nd Stage Seal Press
VibDE	Vibration in DE bearing
VibNDE	Vibration in NDE bearing
Ampere	The current used by the pump

Evaluating VIB DE and VIB NDE bearings is crucial because they support pump shaft alignment, and their condition directly impacts operational efficiency and reliability [7]. This comprehensive compilation serves as a foundational tool for a nuanced examination, allowing for a detailed assessment of the factors that significantly influence the functionality and performance of the WIPs system.

Preprocessing steps include data cleansing, type conversion, outlier detection, feature selection, correlation analysis, and normalization. Data cleansing addresses inconsistencies and missing values. Type conversion ensures uniformity in calculations. Outlier detection manages outliers that could hinder model training. Feature selection, guided by correlation analysis, reduces model complexity. Normalization and scaling ensure uniform feature scales for effective ML.

C. *Data Analysis*

An extensive repository of raw data encompassing operational parameters and cumulative operational hours was initially collected. The dataset refinement process involved identifying salient data points that had a high correlation with WIPs. This process aimed to enhance model interpretability and counteract dimensionality augmentation. The selected key operational parameters are Ampere, VibDE, VibNDE, 1st_Press, and 2nd_Press. Feature selection is approached judiciously, ensuring that only the most influential features are considered for each model. The provided code snippet employs the "sweetviz" Python library for exploratory data analysis (EDA), generating comprehensive reports for informed decision making. The analysis involves data cleaning, exclusion of string entries, and construction of a correlation matrix to assess inter-feature relationships. The Correlation Matrix of Feature Influences helps in selecting the pertinent features shown in Fig. 2, contributing to dimensionality re-duction and enhancing model accuracy. This matrix analysis promotes a nuanced understanding of data structure, confounding factors, and noise, ultimately improving the robustness of the research.

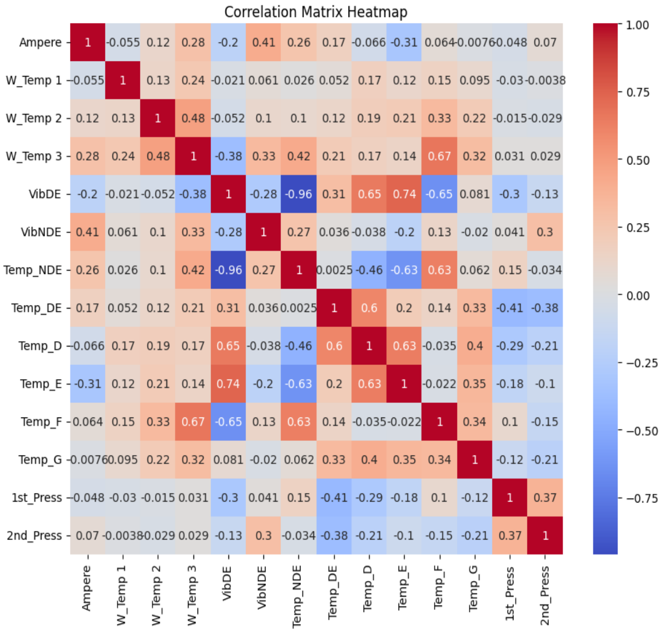

Fig. 2. Correlation matrix of feature influences.

D. *Comparative Modeling of Operational Parameters*

In this study, an assessment is made of two distinct modeling techniques concerning operational parameters. The Light Gradient Boosting Machine (LGBM), eXtreme Gradient Boosting (XGBoost), and random forest methodologies serve as the primary frameworks for examination [22]. The deliberate selection of multiple modeling techniques enriches the empirical rigor of the research, enabling a comparative analysis of their predictive capabilities and generalization capacities. Rooted in the gradient boosting paradigm, LGBM, XGBoost, and random forest models iteratively enhance accuracy by sequentially fitting weak learners to residuals, capturing intricate data patterns [23]. Chosen for their success in various domains, especially tasks requiring high accuracy and efficiency, these models offer a range of hyperparameter configurations for fine-tuning. By employing two distinct models, the aim is to comprehensively understand their performance variances, strengths, and limitations in capturing the intricate interplay among operational parameters. This deliberate and empirically driven choice enhances the scientific rigor of the research, enriches the depth of analysis, and facilitates a nuanced interpretation of the obtained results.

1) XGboost prediction model: The XGBoost model stands out as a powerful ML algorithm deeply rooted in gradient boosting. Recognized for its versatility and exceptional performance, XGBoost is a valuable asset across various data science and ML domains [24]. Operating as an ensemble of decision trees, XGBoost employs a sequential correction approach to iteratively enhance model performance. This methodology allows it to capture intricate relationships within datasets, while integrated tree pruning controls model complexity for improved computational efficiency [22]. Engineered for optimized speed and efficiency, XGBoost is scalable for large datasets through parallel processing. Its adaptability spans diverse data types, and it excels in both regression and classification tasks [25]. With features such as metric-based feature importance, handling imbalanced datasets, and robust regularization techniques, XGBoost emerges as a versatile and powerful algorithm, known for efficiently tackling complex tasks [26].

2) LightGBM prediction model: The LightGBM model is a high-performance open-source software library designed specifically for gradient boosting. Renowned for its speed, resource efficiency, and scalability, LightGBM finds applications in diverse ML tasks such as classification, regression, and ranking [27]. Unlike some gradient-boosting algorithms, LightGBM employs decision trees as base learners, contributing to its exceptional efficiency. Innovative techniques like Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) further enhance its capabilities for reducing model variance and optimizing feature selection. Notably, LightGBM stands out for its remarkable speed, making it one of the fastest ML libraries capable of efficiently training models on extensive datasets. Its efficiency in memory usage optimization allows it to handle datasets surpassing the memory capacity of alternative libraries [28]. Additionally, LightGBM consistently achieves impressive accuracy, delivering state-of-the-art results across a range of ML tasks, further solidifying its reputation as a powerful and scalable library.

3) Random forest prediction model: Random forest regression, a highly regarded technique in ML, is known for its versatility, robustness, and superior predictive accuracy. This ensemble method combines multiple decision trees, excelling in handling complex nonlinear relationships in diverse prediction tasks [29]. Its acclaim stems from consistently outperforming singular decision trees and other regression models by capturing intricate nonlinear relationships while mitigating overfitting. While recognized for its robustness to outliers and noise, a closer examination is crucial to understand its limits [30]. The algorithm handles high-dimensional data well, but scalability and efficiency are key. Feature importance analysis offers insights, but robustness across datasets is crucial [30]. This analysis aims to provide a nuanced view of random forest regression, emphasizing its strengths and offering alternative scenarios.

III. Results

This section compares XGBoost, LGBM, and Random Forest algorithms in predicting parameters like pressure, vibration, and Ampere values, aiming to assess their real-world effectiveness and reliability.

A. *Comparative Analysis of Algorithms*

1) XGBoost model for pressure prediction: insights and visual analysis: Tables II and III present a detailed analysis of the XGBoost model's performance metrics for Seal Pressure 1 and Seal Pressure 2, respectively. The XGBoost model for Seal Pressure 1 exhibits commendable metrics, with an RMSE of 5.61, an Mean Absolute Error (MAE) of 2.38, and an R-squared value of 0.93. These metrics suggest that the model provides accurate predictions, capturing a significant portion of the variance in the data for Seal Pressure 1.

TABLE II. XGBOOST MODEL METRICS FOR SEAL PRESSURE 1

Model	RMSE	MAE	R squared
XGBoost	5.61	2.38	0.93

Table III focuses on Seal Pressure 2, demonstrating the XGBoost model's robust performance with an RMSE of 9.41, an MAE of 4.87, and an R-squared value of 0.91. Despite the slightly higher RMSE and MAE compared to Seal Pressure 1, the model exhibits reasonable accuracy. Interpretation of these metrics should consider specific application needs and tolerance for errors. The R-squared values, nearing 1.0, signify strong correlation, while RMSE and MAE offer insights into prediction errors. Overall, the XGBoost model proves effective in predicting both Seal Pressure 1 and Seal Pressure 2, providing valuable insights for practitioners in pressure prediction scenarios.

TABLE III. XGBOOST MODEL METRICS FOR SEAL PRESSURE 2

Model	RMSE	MAE	R squared
XGBoost	9.41	4.87	0.91

Known for its robustness with complex datasets, the XGBoost algorithm excels in pressure forecasting. This overview explores its application for precise pressure prediction, crucial in diverse sectors. Fig. 3 and Fig. 4 visually compare actual and predicted values, focusing on the initial 25 predictions. Logarithmic scaling enhances clarity, and the limited predictions prevent overcrowding, allowing for a focused evaluation of XGBoost's predictive accuracy.

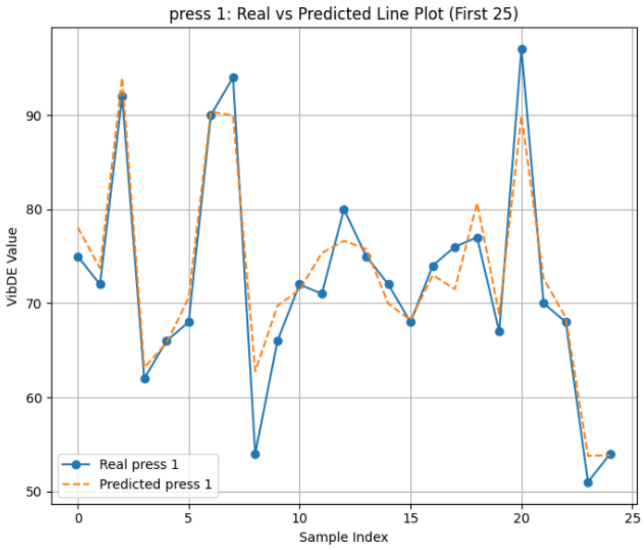

Fig. 3. XGBoost model—line plot analysis of actual vs. predicted pressure 1 (first 25 predictions).

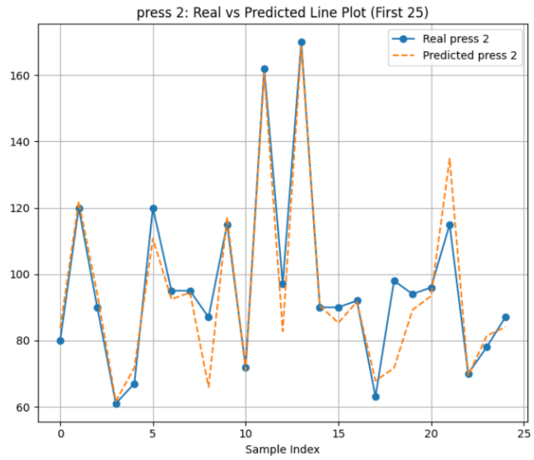

Fig. 4. XGBoost model—focused visualization with limited predictions pressure 2 (first 25 predictions).

2) XGBoost Model for vibration prediction: Insights and visual analysis: Evaluating VIB DE and VIB NDE bearings is crucial for pump system health. In Table IV, focusing on VIB DE, the XGBoost model shows strong performance with an RMSE of 6.32, an MAE of 3.80, and a high R-squared value of 0.99, indicating accurate predictions.

TABLE IV. VIBRATION IN VIB DE—XGBOOST MODEL METRICS

Model	RMSE	MAE	R squared
XGBoost	6.32	3.80	0.99

Table V, evaluating VIB NDE, shows excellence with an RMSE of 3.34, an MAE of 3.80, and an impressive R-squared value of 0.99. These metrics highlight the XGBoost model's proficiency in predicting Non-Drive End-bearing vibration. Comparing Tables IV and V reveals consistent high performance in predicting vibration for both Drive End and Non-Drive End bearings. Strong R-squared values indicate a robust correlation, emphasizing model reliability. Similar MAE values suggest consistent accuracy. Accurate vibration prediction is crucial for identifying potential issues and preventing malfunctions, showcasing the XGBoost model's effectiveness in addressing vibration complexities for overall pump system health and longevity.

TABLE V. VIBRATION IN VIB NDE—XGBOOST MODEL METRICS

Model	RMSE	MAE	R squared
XGBoost	3.34	3.80	0.99

Fig. 5 and Fig. 6 depict the XGBoost model's performance in predicting VIB DE and VIB NDE. Using a line graphs, the visualizations compare the initial 25 predictions with actual values. Applying a logarithmic function enhances scale and normalizes data for a clearer presentation, reducing clutter. The intentional limit of 25 predictions prevents graph overcrowding, ensuring a focused and comprehensive representation.

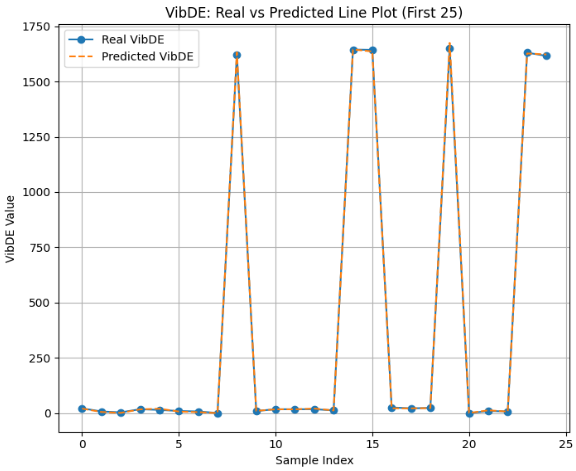

Fig. 5. XGBoost model predictions vs. actual values for VIB DE (line graph).

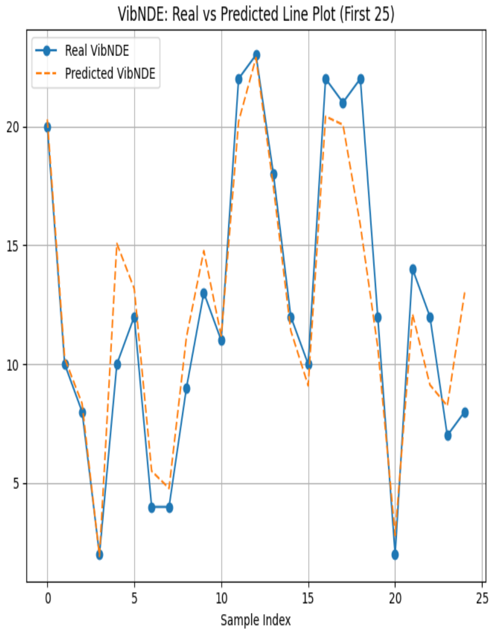

Fig. 6. XGBoost model predictions vs. actual values for VIB NDE (line graph).

3) XGBoost model for AMPERE prediction: insights and visual analysis: Table VI details the predictive performance metrics for the XGBoost model in Ampere measurements forecasting. With an RMSE of 3.28, an MAE of 2.25, and an R-squared value of 0.79, the model shows reasonably accurate predictions for Ampere. While demonstrating proficiency, there is room for improvement, particularly in explaining variance. These metrics serve as benchmarks, guiding potential refinements to enhance precision in future iterations. Insights from Table VI contribute to ongoing efforts to optimize parameters or explore alternative methodologies, crucial for fine-tuning the model and maximizing its effectiveness in real-world applications where accurate Ampere predictions are crucial.

TABLE VI. XGBOOST MODEL METRICS FOR AMPERE PREDICTION

Model	RMSE	MAE	R squared
XGBoost	3.28	2.25	0.79

Fig. 7 visually analyze the XGBoost model's predictive performance, comparing the first 25 predictions with actual values using grouped line plots. The contrast between actual and predicted values highlights the model's accuracy, with line plots depicting continuous performance trends. Applying a logarithmic function enhances clarity and comparability, ensuring better scale and data normalization for a coherent representation. Focusing on the initial 25 predictions prevents visual overcrowding, thereby facilitating a detailed examination of early-phase accuracy.

This approach allows for effective pattern recognition and insights. Overall, these visualizations offer a comprehensive and accessible assessment of the XGBoost model's predictive capabilities, leveraging a combination of graphs and thoughtful data transformations for nuanced understanding and valuable insight.

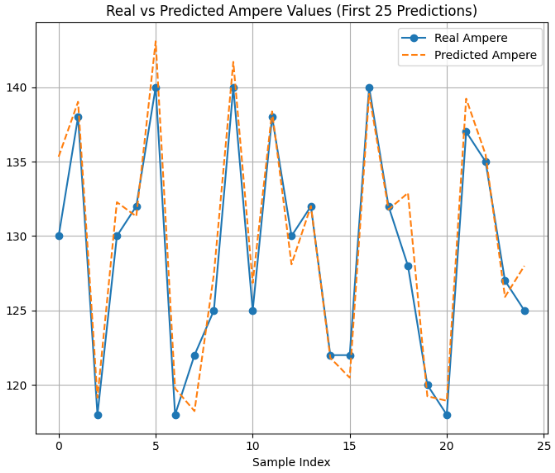

Fig. 7. XGBoost model—line plot analysis of actual vs. predicted values (first 25 predictions).

4) LGBM model for pressure prediction: insights and visual analysis: Examining the outcomes presented in Table VII, it is evident that the LGBM model achieves noteworthy metrics for Seal Pressure 1, with an RMSE of 5.29, an MAE of 2.22, and an R-squared value of 0.94.

TABLE VII. LGBM MODEL METRICS FOR SEAL PRESSURE 1

Model	RMSE	MAE	R squared
LGBM	5.29	2.22	0.94

Moving to Table VIII, the LGBM model's performance for Seal Pressure 2 is notable, featuring an RMSE of 8.98, an MAE of 5.17, and an R-squared value of 0.91. These tabulated results provide a detailed overview of the LGBM model's effectiveness for both Seal Pressure 1 and Seal Pressure 2, facilitating a comprehensive evaluation of its predictive capabilities.

TABLE VIII. LGBM MODEL METRICS FOR SEAL PRESSURE 2

Model	RMSE	MAE	R squared
LGBM	8.98	5.17	0.91

Fig. 8 and Fig. 9 visually showcase the LGBM model's predictive performance in pressure forecasting. Utilizing line plots for a comparative analysis of the initial 25 predictions against actual values, the visualization highlights the LGBM model's effectiveness. Applying a logarithmic function enhances clarity and maintains a normalized scale, ensuring a clearer and less cluttered representation. Focusing on the initial 25 predictions allows for detailed examination of early accuracy, preventing graph overcrowding, and allowing for a interpretable and focused visual representation. These visualizations offer valuable insights into the LGBM model's predictive capabilities, aiding in assessing its performance and reliability in pressure prediction scenarios. The graphical representation facilitates an intuitive understanding of the model's behavior, contributing to a comprehensive evaluation.

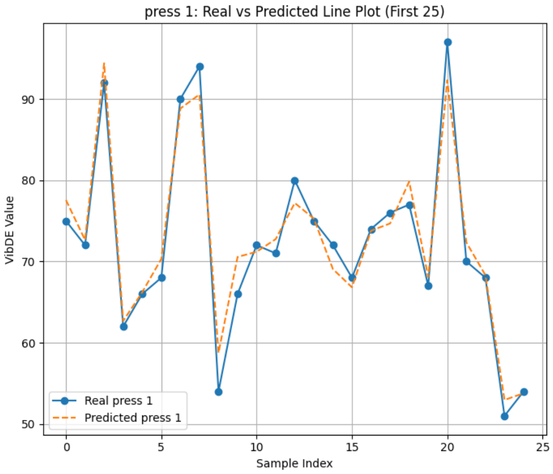

Fig. 8. LGBM Model—Line Plot Analysis of Actual vs. Predicted Pressure (First 25 Predictions).

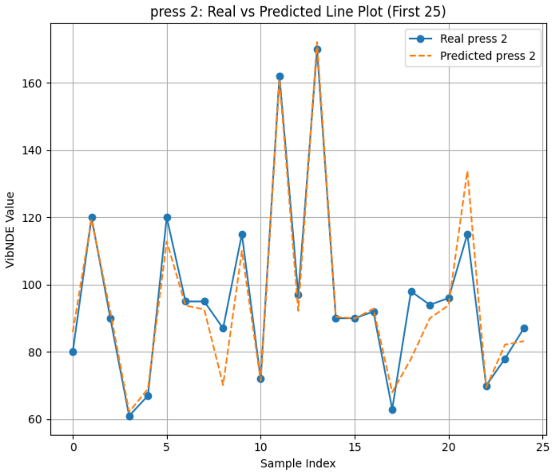

Fig. 9. LGBM Model—Focused Visualization with Limited Predictions (First 25 Predictions).

5) LGBM model for vibration prediction: Insights and visual analysis: The visualizations presented in Fig. 10 depict the performance of the LGBM model in predicting VIB DE and VIB NDE values. The line graphs facilitate a comprehensive comparison of the first 25 predictions with their corresponding actual values. Applying the log function to the values serves the purpose of achieving a better scale and normalization, leading to a clearer and less cluttered visualization. This transformation enhances the interpretability of the data, making it easier to discern patterns and trends in the model's predictions. By limiting the display to the first 25 predictions, the graphs avoid overcrowding, allowing for a focused examination of the model's accuracy in capturing the actual values. This selective approach aids in identifying any discrepancies or areas where the model may exhibit strengths or weaknesses. The discussion of these visualizations should involve a detailed analysis of how well the LGBM model aligns with the actual values, considering factors such as precision, accuracy, and potential areas for improvement. Additionally, any notable patterns or deviations between predicted and actual values should be highlighted and discussed to provide insights into the model's performance and its applicability in predicting VIB DE and VIB NDE.

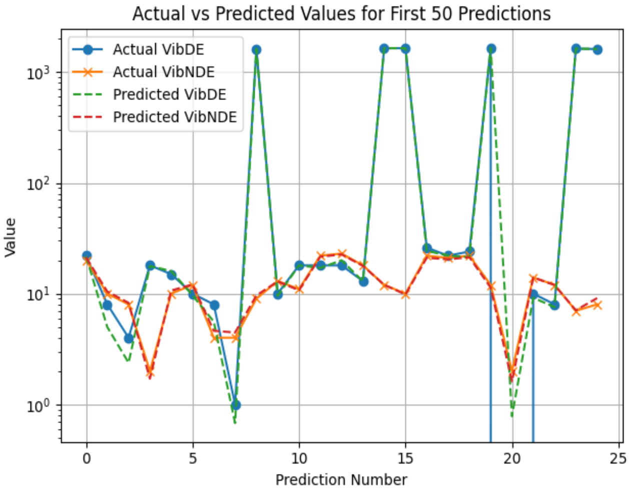

Fig. 10. LGBM Model Predictions vs. Actual Values for VIB NDE.

The LGBM model for VIB DE demonstrates exceptional performance, as indicated by the metrics in Table IX. The Root Mean Squared Error (RMSE) stands impressively low at 45.22, signifying minimal deviation between predicted and actual values. Complementing this, the MAE is commendably low at 6.07, reinforcing the model's accuracy. The high R-squared value of 0.99 emphasizes an excellent fit, showcasing the model's ability to explain the variance in VIB DE.

TABLE IX. PERFORMANCE METRICS FOR VIB DE USING LGBM

Model	RMSE	MAE	R squared
LGBM	45.22	6.07	0.99

Turning to VIB NDE, the LGBM model continues to exhibit strong predictive capabilities, as highlighted in Table X. The RMSE is notably low at 3.46, indicating minimal prediction errors. The MAE, standing at 2.32, further emphasizes the accuracy of the model, with low absolute differences between predicted and actual values. While the R-squared value of 0.84 is slightly lower than in VIB DE, it still signifies a robust model fit and reliable predictions for VIB NDE. The Tables IX and X are collectively underscore the effectiveness of the LGBM model in predicting both VIB DE and VIB NDE.

TABLE X. PERFORMANCE METRICS FOR VIB NDE USING LGBM

Model	RMSE	MAE	R squared
LGBM	3.46	2.32	0.84

6) LGBM model for AmperE prediction: Insights and visual Analysis : Fig. 11 offers a detailed analysis of the LGBM model's Ampere prediction performance using grouped line plots. Comparing the initial 25 predictions with actual values, these visualizations provide insights into the model's accuracy. Line plots offer a continuous overview of the model's performance. Applying a logarithmic function enhances interpretability and comparability, ensuring a clearer and less cluttered visualization. Focusing on the first 25 predictions prevents visual congestion, allowing a detailed examination of early-stage accuracy and facilitating the discernment of performance patterns. The line plots thoughtful data transformations contributes to a nuanced understanding of the LGBM model's predictive performance, facilitating valuable insights from the visualized data.

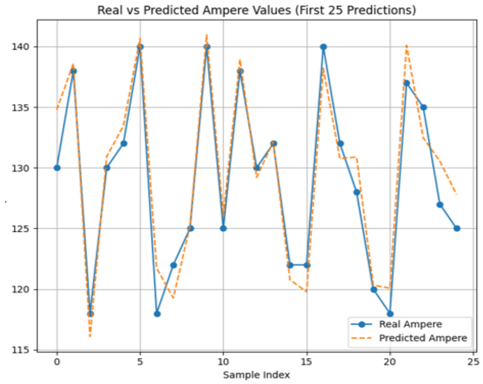

Fig. 11. LGBM Model—Comparison of Actual vs. Predicted Ampere Values (First 25 Predictions).

Table XI succinctly evaluates the LGBM model's performance in predicting Ampere values through the three key metrics—RMSE, MAE, and R-squared. With an RMSE of 3.08 indicating average error magnitude, a low MAE of 2.26 signifying precise predictions, and a high R-squared value of 0.82 showcasing substantial explanatory power, the LGBM model proves effective for Ampere prediction. These metrics affirm the model's reliability and accuracy, positioning it as a valuable tool for forecasting Ampere values across applications. The table provides a concise summary, offering numerical indicators for researchers, practitioners, and decision makers seeking insights into the model's efficacy.

TABLE XI. LGBM MODEL METRICS FOR AMPERE PREDICTION

Model	RMSE	MAE	R squared
LGBM	3.08	2.26	0.82

7) Random forest model for pressure prediction: Insights and visual analysis: Fig. 12 and Fig. 13 visually depict the Random Forest model's predictive performance using line plots, contrasting the initial 25 predictions with actual values. The combination of these visual elements highlights the model's efficacy and provides a comprehensive analysis of its accuracy in predicting the first 25 observations. Applying a logarithmic function enhances clarity and maintains a normalized scale, resulting in a distinct and less cluttered representation. Focusing on the initial 25 predictions offers valuable insights into the model's early predictive behavior, facilitating a detailed examination of accuracy without overcrowding the graphical representation. These visualizations contribute to a nuanced understanding of the Random Forest model's predictive capabilities, offering unique insights through line plots. This visual exploration is crucial for assessing the model's reliability and effectiveness, particularly in scenarios where clarity and precision are paramount.

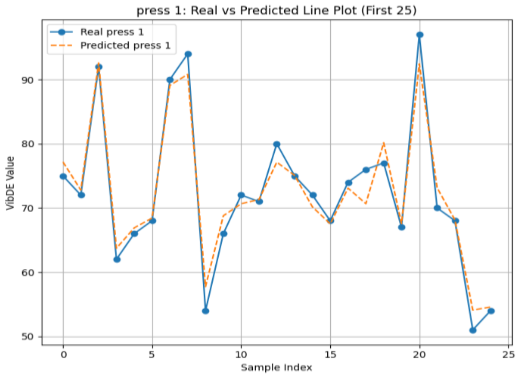

Fig. 12. Random Forest Model—Logarithmic Transformation (First 25 Predictions).

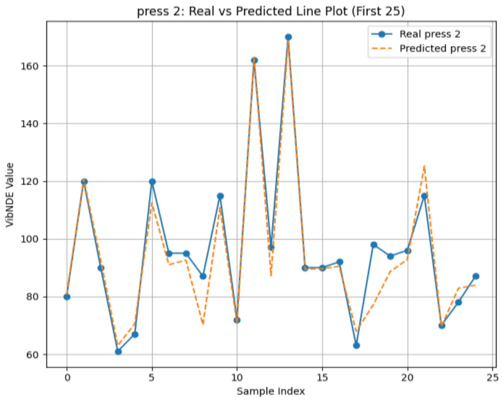

Fig. 13. Random Forest Model—Focused Visualization with Limited Predictions (First 25 Predictions).

In Table XII, the Random Forest model exhibits commendable metrics for Seal Pressure 1, with an RMSE of 4.86, an MAE of 2.01, and an R-squared value of 0.95. These results suggest that the Random Forest model provides accurate predictions, capturing a significant portion of the variance in the data for Seal Pressure 1.

TABLE XII. RANDOM FOREST MODEL METRICS FOR SEAL PRESSURE 1

Model	RMSE	MAE	R squared
Random Forest	4.86	2.01	0.95

Table XIII focuses on Seal Pressure 2, where the Random Forest model demonstrates robust performance with an RMSE of 9.24, an MAE of 4.49, and an R-squared value of 0.91. Despite slightly higher RMSE and MAE values compared to Seal Pressure 1, the model maintains reasonable accuracy. Combined with visualizations, it is evident that the Random Forest model consistently performs well in both seal pressure scenarios. Strong correlation, as indicated by high R-squared values, underscores the model's reliability in predicting seal pressures. The integration of visual and quantitative assessments provides a comprehensive understanding of the Random Forest model's effectiveness.

TABLE XIII. RANDOM FOREST MODEL METRICS FOR SEAL PRESSURE 2

Model	RMSE	MAE	R squared
Random Forest	9.24	4.49	0.91

8) Random forest model for vibration prediction: insights and visual analysis: Fig. 14 and 15 visually analyze the Random Forest model's performance in predicting VIB DE and VIB NDE, employing both bar and line formats to compare the initial 25 predictions with actual values. Applying a log function to the values enhances scaling and normalization, improving visualization clarity. Focusing on the first 25 predictions prevents overcrowding, allowing for a detailed examination of the model's accuracy and nuances in data capture. This approach provides a concise yet meaningful snapshot of the Random Forest model's alignment with actual values, offering valuable insights for further analysis and model refinement.

In Tables XIV and XV, the performance metrics of the Random Forest model in predicting VIB DE and VIB NDE are presented. In Table XIV, the model achieved an RMSE of 9.53, indicating precise predictions with a low average magnitude of errors. The MAE of 3.80 further confirms the model's accuracy, as it represents the average absolute difference between predicted and actual values. The exceptionally high R squared value of 0.99 signifies a remarkable goodness of fit, explaining 99% of the variance in VIB DE.

Turning to Table XV, the Random Forest model showcased a notable RMSE of 3.58 for VIB NDE, indicating accurate predictions with a low average magnitude of errors. The MAE of 2.16 reinforces the model's reliability, showcasing a relatively small average absolute difference from the actual values. The R squared value of 0.83 highlights a strong fit, accounting for 83% of the variance in VIB NDE.

In summary, the Random Forest model demonstrates exceptional predictive accuracy for both VIB DE and VIB NDE. The low RMSE and MAE values, coupled with high R-squared values, underscore the model's effectiveness in capturing and explaining the variance in vibration data. These findings affirm the Random Forest model as a robust tool for vibration prediction in both conditions.

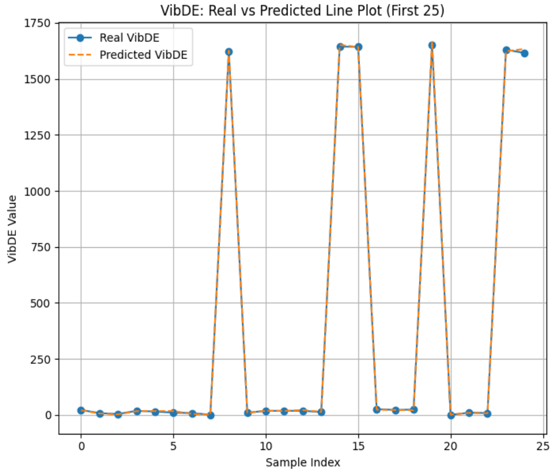

Fig. 14. RandomForest_VIB_DE_Performance_Line.

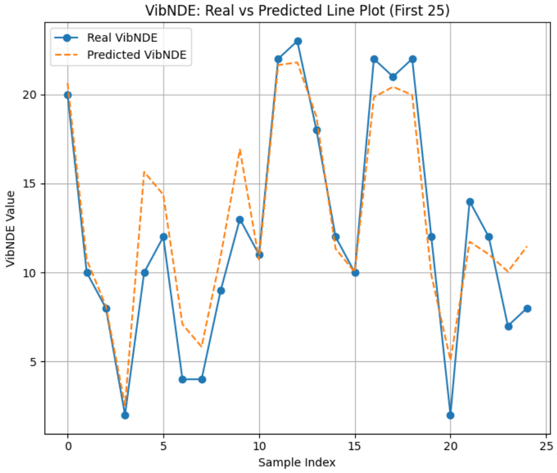

Fig. 15. RandomForest_VIB_NDE_Performance_Line.

TABLE XIV. RANDOM FOREST VIB DE PERFORMANCE

Model	RMSE	MAE	R squared
Random Forest	9.53	3.80	0.99

TABLE XV. RANDOM FOREST VIB NDE PERFORMANCE

Model	RMSE	MAE	R squared
Random Forest	3.58	2.16	0.83

9) Random forest model ampere prediction-insights and visual analysis: The visual representation in Fig. 16 offers a detailed examination of the Random Forest model's efficacy in predicting Ampere values. Utilizing line plots, these visualizations facilitate a comprehensive comparison of the initial 25 predictions made by the Random Forest model against the actual values. Applying a logarithmic function enhances interpretability and normalization of the data, resulting in a clearer representation of the Random Forest model's predictive accuracy. Focusing on the first 25 predictions prevents visual congestion, allowing a detailed examination of early-stage accuracy and providing valuable insights into the model's predictive behavior for Ampere values.

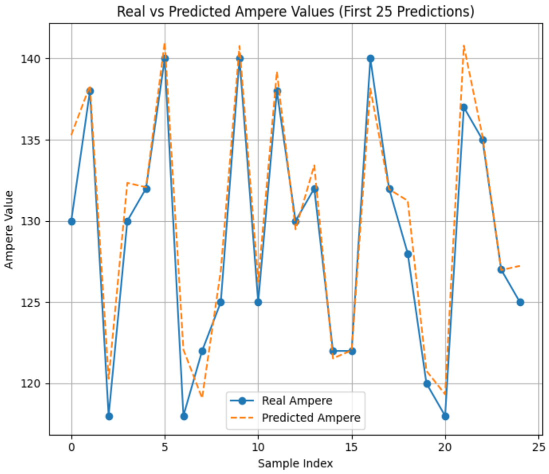

Fig. 16. Random Forest Model - Line Plot Analysis of Actual vs. Predicted Ampere Values (First 25 Predictions).

The performance metrics for Ampere prediction using the Random Forest model, as presented in Table 16, showcase its commendable accuracy. With a low RMSE of 2.96 and a close match between predicted and actual values indicated by an MAE of 2.10, the model demonstrates robust predictive capabilities. The high R-squared value of 0.83 further emphasizes its ability to explain a substantial proportion of the variance in Ampere measurements. In summary, these metrics affirm the Random Forest model's effectiveness in delivering accurate Ampere predictions, underscoring its potential for reliable forecasting in relevant applications.

TABLE XVI. RANDOM FOREST MODEL METRICS FOR AMPERE PREDICTION

Model	RMSE	MAE	R squared
Random Forest	2.96	2.10	0.83

B. *Ensemble Approach for Maintenance Prediction*

Ensembling XGBoost, LightGBM, and Random Forest models in maintenance prediction, as shown in Fig. 17, combines their predictive outputs to boost accuracy and robustness. This approach capitalizes on the unique strengths of each model—XGBoost's efficiency, LightGBM's speed, and Random Forest's robustness. By fusing these models, the ensemble leverages their collective intelligence, mitigating individual weaknesses. The visual in Fig. 17 illustrates how this collaboration forms a cohesive and reliable maintenance prediction framework, enhancing resilience to uncertainties and improving overall effectiveness in complex scenarios.

To improve predictive outcomes, an ensembling approach combines three models—XGBoost, LGBM, and Random Forest. Ensemble learning—leveraging the strengths of multiple models, mitigating risks of overfitting and underfitting, and enhancing predictive accuracy [17]. It addresses common ML challenges, making models more robust and stable. Ensemble learning excels in capturing complex data relationships, proving effective in scenarios where a single model may struggle. In the voting mechanism, soft voting is preferred for its regression nature. Soft voting, averaging predicted probabilities, offers a nuanced approach over hard voting, which is particularly beneficial when models exhibit varying confidence levels. The advantages of soft voting include enhanced predictive performance and flexibility, making it suitable for imbalanced datasets. In conclusion, strategic ensembling—especially through soft voting—stands as a simple yet effective method for improving ML model performance, particularly in scenarios with varying confidence levels.

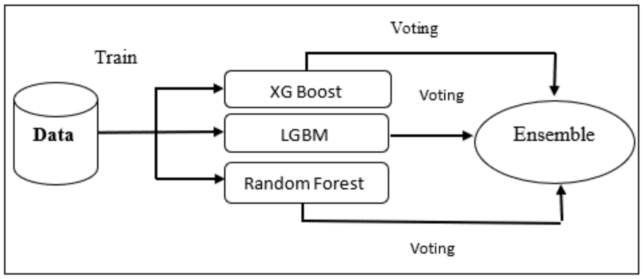

Fig. 17. Ensemble of XGBoost, LightGBM, and Random Forest Models for Maintenance Prediction.

1) Results and analysis of ensemble model prediction for seal pressure: The results from the evaluation of the pump's seal pressure models, presented in Tables XVII and XVIII, demonstrate the performance of XGBoost, LGBM, Random Forest, and the ensemble approach. For Seal Pressure 1 (Table XVII), the standalone models perform well: XGBoost (RMSE: 5.61, MAE: 2.38, R-squared: 0.93); LGBM (RMSE: 5.29, MAE: 2.22, R-squared: 0.94); and Random Forest (RMSE: 4.86, MAE: 2.01, R-squared: 0.95). The Ensemble model significantly outperforms these, with RMSE: 1.94, MAE: 0.92, and R-squared: 0.99. Similarly, for Seal Pressure 2 (Table XVII), the standalone models show strong performance: XGBoost (RMSE: 9.41, MAE: 4.87, R-squared: 0.91); LGBM (RMSE: 8.98, MAE: 5.17, R-squared: 0.91); and Random Forest (RMSE: 9.24, MAE: 4.49, R-squared: 0.91). Again, the Ensemble model achieves superior results, with RMSE: 2.94, MAE: 1.74, and R-squared: 0.99. These findings underscore the Ensemble model's enhanced predictive capabilities for both Seal Pressure 1 and Seal Pressure 2. The marked improvement in RMSE, MAE, and R-squared values indicates that the ensemble approach effectively combines the strengths of XGBoost, LGBM, and Random Forest, leading to higher accuracy and reduced prediction errors.

TABLE XVII. INDIVIDUAL AND ENSEMBLE MODEL METRICS FOR SEAL PRESSURE 1

Model	RMSE	MAE	R squared
XGBoost	5.61	2.38	0.93
LGBM	5.29	2.22	0.94
Random Forest	4.86	2.01	0.95
Ensemble	1.94	0.92	0.99

Table XVIII provides an insightful comparison of metrics for Seal Pressure 2. The standalone models (XGBoost, LGBM, Random Forest) exhibit respectable performance, while the Ensemble model consistently outshines them across all metrics. This emphasizes the Ensemble model’s effectiveness in enhancing accuracy and reducing errors in predicting Seal Pressure 2.

TABLE XVIII. INDIVIDUAL AND ENSEMBLE MODEL METRICS FOR SEAL PRESSURE 2

Model	RMSE	MAE	R squared
XGBoost	9.41	4.87	0.91
LGBM	8.98	5.17	0.91
Random Forest	9.24	4.49	0.91
Ensemble	2.94	1.74	0.99

From Tables XVII and XVIII, it is evident that, for both Seal Pressures 1 and 2, the error rate is notably higher when employing stand-alone models compared to the Ensemble model. Utilizing three distinct error metrics—root mean squared, mean average precision, and error squared—we consistently observe that the Ensemble model outperforms each of the stand-alone models. This observation underscores the potential of combining multiple models into an ensemble, demonstrating its efficacy in achieving heightened accuracy and minimizing prediction errors across diverse evaluation metrics. Fig. 18 offers a visual representation of the Ensemble model's performance in predicting Seal Pressure 1, employing line plots to compare the initial 25 predictions with their actual values. The visual presentation underscores the effectiveness of the Ensemble model through line plots, providing a comprehensive analysis of its accuracy in predicting the first 25 observations of Seal Pressure 1. The application of a logarithmic function to the values serves the dual purpose of enhancing the visualization's clarity and maintaining a normalized scale. This transformation results in a more distinct and less cluttered representation of the model's performance. Focusing on the initial 25 predictions ensures a detailed examination of the Ensemble model's accuracy during the initial phase, contributing to a nuanced understanding of its predictive behavior. This approach also helps to prevent overcrowding in the graphs, ensuring that the visual representation remains interpretable and focused. These visualizations, in conjunction with quantitative metrics, contribute to a comprehensive evaluation of the Ensemble model's reliability and effectiveness in predicting Seal Pressure 1. The integration of visual and quantitative assessments enhances the overall understanding of the model's predictive capabilities and aids in decision making for real-world applications.

Fig. 19 explores the Ensemble model's performance in predicting Seal Pressure 2, comparing the initial 25 predictions with actual values using line plots. The visual representation highlights the model's accuracy, providing a comprehensive analysis of its performance. Applying a logarithmic function enhances clarity and maintains a normalized scale, resulting in a clearer and more distinct representation. Focusing on the initial 25 predictions allows for a detailed examination of the model's accuracy during the early phase, contributing to a nuanced understanding of its predictive behavior. This approach ensures interpretability by preventing graph overcrowding. The visualizations, complemented by quantitative metrics, offer a thorough evaluation of the Ensemble model's reliability and effectiveness in predicting Seal Pressure 2, enhancing the overall understanding of its predictive capabilities for real-world applications.

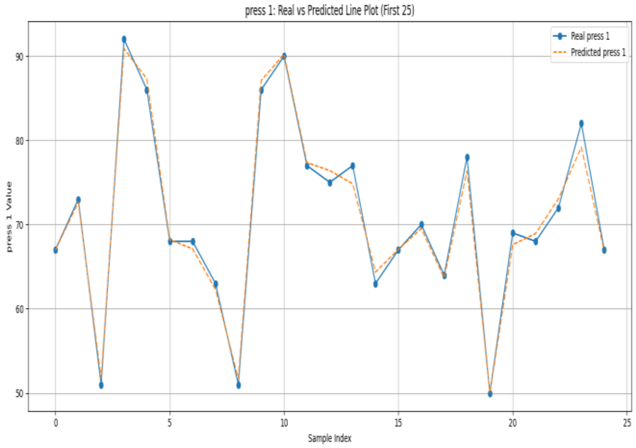

Fig. 18. Ensemble Model—Line Plot Analysis of Actual vs. Predicted Seal Pressure 1 (First 25 Predictions).

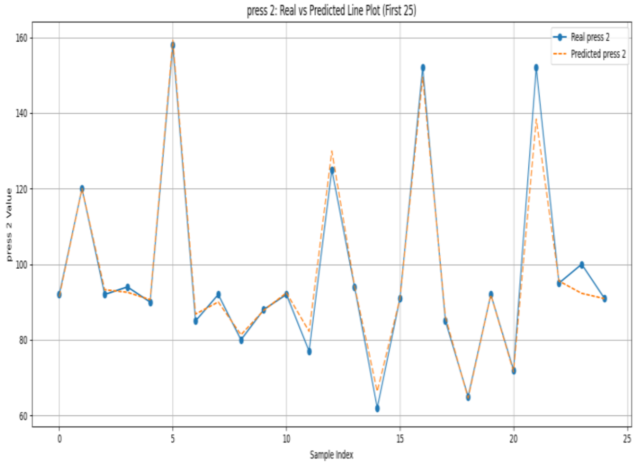

Fig. 19. Ensemble Model—Line Plot Analysis of Actual vs. Predicted Seal Pressure 2.

The examination of both the tables and graph reveals a noticeable elevation in error metrics for individual models. However, a compelling contrast emerges when these models are amalgamated, resulting in a substantial reduction in error, as evidenced by the graphical representations. This improvement is distinctly illustrated by the diminished variance between the actual and predicted lines in the line graph. A comparative analysis of the line graphs for the Ensemble model's underscores the significant enhancement achieved by combining individual models, resulting in a more accurate and reliable predictive outcome.

2) Results and analysis of ensemble model prediction for vibration prediction: The evaluation of the Vibration Diagnoses Equipment (VIB DE) and Vibration Non-Diagnoses Equipment (VIB NDE) models provides valuable insights into their predictive performance for bearing vibrations using RMSE, MAE, and R-squared metrics. In Table XIX, the VIB DE models show high RMSE values, indicating substantial prediction errors, with the LGBM model having an exceptionally high RMSE of 45.22. However, the Ensemble model reduces the RMSE to 8.60, demonstrating the effectiveness of combining the models. The decline in MAE and consistently high R-squared values further emphasize the Ensemble model's superior ability to reduce prediction errors and improve accuracy.

TABLE XIX. INDIVIDUAL AND ENSEMBLE MODEL METRICS FOR VIB DE MODEL METRICS

Model	RMSE	MAE	R squared
XGBoost	6.32	3.80	0.99
LGBM	45.22	6.07	0.99
Random Forest	9.53	3.80	0.99
Ensemble	8.60	1.42	0.99

In Table XX, individual models such as LGBM and Random Forest show notable RMSE values for VIB NDE evaluation. The Ensemble model significantly enhances accuracy with lower RMSE and MAE and higher R-squared values. Interestingly, XGBoost slightly outperforms the Ensemble for VIB DE, indicating algorithmic influence. Graphical and tabular presentations consistently illustrate the Ensemble model's superior performance in error metrics, emphasizing its effectiveness in predicting and managing vibration issues in critical system maintenance.

TABLE XX. INDIVIDUAL AND ENSEMBLE MODEL METRICS FOR VIB NDE MODEL METRICS

Model	RMSE	MAE	R squared
XGBoost	3.34	3.80	0.99
LGBM	3.46	2.32	0.84
Random Forest	3.58	2.16	0.83
Ensemble	1.17	0.75	0.98

Fig. 20 showcases the ensemble of XGBoost, LightGBM, and Random Forest models, leveraging their strengths for accurate and robust maintenance predictions. By integrating diverse perspectives, the ensemble enhances resilience to uncertainties, offering a comprehensive solution for complex operational scenarios. This collaborative approach mitigates individual weaknesses, leading to improved performance compared to standalone models and enhancing the effectiveness of maintenance prediction systems.

Examining the table and graphs reveals notable error metrics for individual models, indicating relatively high errors. However, a significant improvement is evident when these models are integrated into an ensemble. The comparison of actual and predicted lines in the line graph vividly illustrates this enhancement. The Ensemble model, depicted in the graphs, showcases a considerable reduction in error compared to individual models.

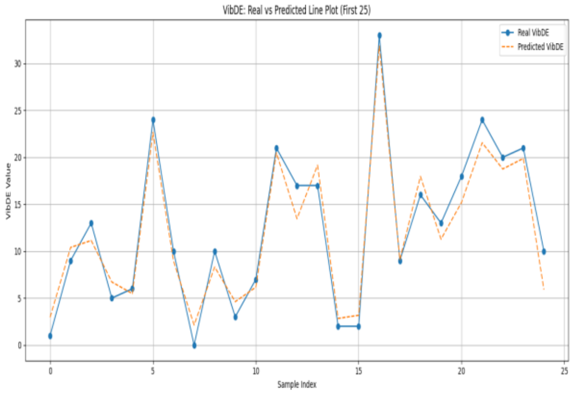

Fig. 1. Ensemble Model's Performance for VIB NDE (Line Graphs).

3) Results and analysis of ensemble model prediction for ampere prediction: When evaluating power consumption, Ampere readings serve as a crucial parameter, with higher readings indicative of increased power consumption. In scenarios where direct current measurements are unavailable, employing ML models becomes a viable option for prediction. In this context, three distinct models—XGBoost, LGBM, and Random Forest—were implemented to forecast Ampere readings. Subsequently, these models were amalgamated into an Ensemble model to harness their collective predictive capabilities. The performance metrics, including RMSE, mean average error (MAE), and R-squared values, were employed to assess the accuracy of each model and the Ensemble model's. The results are presented in Table XXI.

TABLE XXI. AMPERE PREDICTION MODEL PERFORMANCE METRICS

Model	RMSE	MAE	R squared
XGBoost	3.28	2.25	0.79
LGBM	3.08	2.26	0.82
Random Forest	2.96	2.10	0.83
Ensemble	1.69	0.94	0.95

Analysis of the table indicates a substantial decrease in error rates when utilizing the Ensemble model as opposed to individual models. Employing three distinct error metrics consistently demonstrates the superior performance of the Ensemble model, affirming its effectiveness in achieving heightened accuracy and minimizing prediction errors. This underscores the value of combining diverse models to enhance predictive capabilities. The visual representation above showcases the performance of the Ensemble model for the Ampere prediction. Utilizing line graph in Fig. 21, the graph allows a comparison of the first 25 predictions with their actual values. To enhance clarity and minimize clutter, a logarithmic function has been applied to the values, resulting in a more normalized and visually accessible presentation. The decision to focus on 25 predictions ensures a concise and uncluttered depiction, facilitating a clear understanding of the Ensemble model's effectiveness in predicting Ampere values.

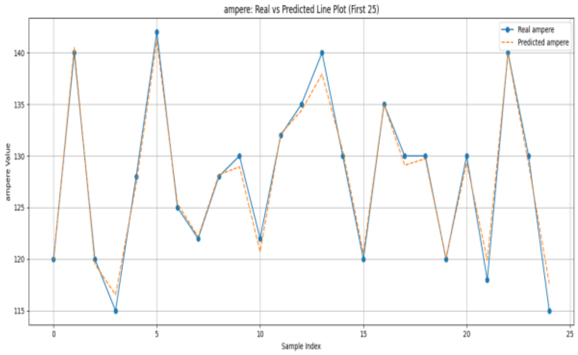

Fig. 2. Line Graph for Ampere Ensemble Model Performance.

Fig. 21 illustrates the performance of the Ensemble model for the Ampere prediction. These visualizations showcase a comparison between the first 25 predictions and their actual values, employing both bar and line graphs. To enhance clarity and maintain a cleaner display, a logarithmic function has been applied to the values, ensuring a more balanced scale. The limited focus on 25 predictions prevents graph overcrowding, facilitating a more straightforward interpretation of the results.

IV. Discussion

This study presents a novel approach to predictive maintenance (PdM) in the oil and gas industry by utilizing an ensemble of three machine learning algorithms—XGBoost, Light Gradient-Boosting Machine (LGBM), and Random Forest. The ensemble model consistently outperformed individual models, demonstrating superior accuracy and reliability across all operational parameters (Seal Pressure 1, Seal Pressure 2, VIB DE, VIB NDE, and Ampere). Notably, the Ensemble model showed lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values and higher R-squared coefficients, indicating better performance in predicting the health of water injection pumps (WIPs). This finding aligns with research conducted by Zhang et al. [1], where combining multiple machine learning models improved prediction accuracy in industrial applications. The novelty of this study lies in the hybridization of these models to predict diverse operational parameters, which is an improvement over the standalone approaches often used in previous research. For example, Janssens et al. [2] applied Convolutional Neural Networks (CNNs) to detect anomalies using infrared thermal images, but the study was limited in its ability to handle different operational variables. Similarly, Sampaio et al. [3] explored Artificial Neural Networks (ANNs) for motor failure prediction but did not consider the comprehensive evaluation of various predictive models. The Ensemble model, in contrast, demonstrates a broader applicability by leveraging the strengths of multiple algorithms, thus mitigating the individual limitations of each method and achieving more robust and reliable predictions. Furthermore, the results of this study highlight the importance of addressing operational parameters beyond vibration data, such as seal pressure and amperage, which are often overlooked in predictive maintenance studies. This is in contrast to previous studies that focused primarily on vibration data for fault detection and failure prediction [4], [5]. By including additional operational parameters, this research provides a more holistic approach to predictive maintenance, which could lead to more effective early detection of potential failures in WIPs, ultimately reducing unplanned downtime and improving operational efficiency. The ensemble approach presented in this study also reflects a growing trend in the literature toward integrating multiple machine learning techniques to enhance the accuracy of PdM models. This is consistent with the findings of Kim et al. [6], who demonstrated that ensemble methods, when applied to predictive maintenance in industrial equipment, resulted in significant improvements in both prediction accuracy and reliability. The success of the ensemble model in this study suggests that further refinement and adaptation of this methodology could be applied to other critical equipment within the oil and gas industry, as well as other industrial sectors facing similar challenges with unplanned downtime. In sum, the findings of this study contribute to the growing body of research on predictive maintenance by offering an innovative methodology for improving the accuracy of failure predictions in water injection pumps. The ensemble model's superior performance, when compared to individual models, underscores the potential for more accurate and reliable predictive systems in industrial applications. The results also pave the way for further exploration of hybrid machine learning models in PdM, particularly in industries with complex operational environments, such as oil and gas.

V. Limitations

The Ensemble model, comprising XGBoost, LGBM, and Random Forest, demonstrates commendable predictive accuracy across diverse operational parameters. However, it is imperative to acknowledge certain limitations inherent in the approach. First, the efficacy of the Ensemble model is highly contingent on the quality of the input data. Instances of data inconsistencies, inaccuracies, or a lack of representativeness concerning diverse operating conditions can potentially compromise the performance. Moreover, the representativity of the training set plays a pivotal role. The Ensemble model relies on a training dataset that effectively captures the various operating scenarios of the pump system. Notably, the study focused on predicting pressure, vibration, and amperage for temperature, while considering other features that could contribute to PdM in WIPs within the OGI. These limitations underscore the importance of meticulous data quality assurance, comprehensive representation in training datasets, and ongoing refinement of hyperparameter configurations for the reliable and robust application of the Ensemble model in PdM scenarios.

VI. Future Research Direction

The Ensemble model, comprising XGBoost, LGBM, and Random Forest, consistently demonstrates notable predictive accuracy for various operational parameters in WIPs. However, the model's performance is contingent on high-quality input data and a representative training set. Notably, the sensitivity to hyperparameter configurations requires ongoing optimization efforts. Future research directions could explore advanced ensemble techniques beyond the current models and dynamic hyperparameter tuning mechanisms for autonomous adaptation. Investigating the impact of external factors, transitioning to real-time predictions, enhancing explainability, scalability testing, and integrating the model into existing maintenance systems are promising avenues. Seeking feedback from industry practitioners is vital for refining the model's real-world applicability.

VII. Conclusions

In conclusion, this study demonstrates the effectiveness of combining multiple machine learning algorithms—XGBoost, LGBM, and Random Forest—into an Ensemble model for predictive maintenance of water injection pumps. The Ensemble model consistently outperforms individual algorithms, showcasing superior accuracy through lower RMSE and MAE values, as well as higher R-squared coefficients. By integrating the strengths of these algorithms, the Ensemble model mitigates the limitations of standalone models, offering a more robust and reliable predictive maintenance tool. The results underscore the potential for improving operational efficiency and reducing unplanned downtime in the oil and gas industry. This research not only advances predictive modeling techniques but also highlights the significant implications for enhancing maintenance strategies in industrial applications, ensuring better asset management and cost-effectiveness in critical systems.

References

[1] Gupta, D.; Shah, M. A comprehensive study on artificial intelligence in oil and gas sector. Environ. Sci. Pollut. Res. 2022, 29, 50984–50997.

[2] Trevathan, M.M.T. The Evolution, Not Revolution, of Digital Integration in Oil and Gas. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2020

[3] Almazrouei, S.; Dweiri, F.; Aydin, R.; Alnaqbi, A. A review on the advancements and challenges of artificial intelligence based models for predictive maintenance of water injection pumps in the oil and gas industry. SN Appl. Sci. 2023, 5, 391.

[4] Chen, C.; Fu, H.; Zheng, Y.; Tao, F.; Liu, Y. The advance of digital twin for predictive maintenance: The role and function of machine learning. J. Manuf. Syst. 2023, 71, 581–594.

[5] Ngu, K.M.; Philip, N.; Sahlan, S. Proactive and predictive maintenance strategies and application for instrumentation & control in oil & gas industry. Int. J. Integr. Eng. 2019, 11, 119–130.

[6] Manchadi, O., Ben-Bouazza, F.E., & Jioudi, B. Predictive Maintenance in healthcare system: A Survey. IEEE Access 2023, 11, 61313–61330

[7] Omri, F.; Choura, O.; Taieb, L.H.; Elaoud, S. Prediction of Bearing Fault Effect on the Hydraulic Performances of a Centrifugal Water Pump. J. Vib. Eng. Technol. 2022, 10, 1905–1915.

[8] Xiang, C.; Li, B. Research on ship intelligent manufacturing data monitoring and quality control system based on industrial Internet of Things. Int. J. Adv. Manuf. Technol. 2020, 107, 983–992.

[9] Hornyák, O. The Role of Condition-Based Maintenance in Minimizing Operational Costs. Prod. Syst. Inf. Eng. 2023, 11, 43–53.

[10] Aissani, N.; Beldjilali, B.; Trentesaux, D. Dynamic scheduling of maintenance tasks in the petroleum industry: A reinforcement approach. Eng. Appl. Artif. Intell. 2009, 22, 1089–1103.

[11] Compare, M.; Baraldi, P.; Zio, E. Challenges to IoT-enabled predictive maintenance for industry 4.0. IEEE Internet Things J. 2019, 7, 4585–4597.

[12] Saputelli, L.; Palacios, C.; Bravo, C. Case Studies Involving Machine Learning for Predictive Maintenance in Oil and Gas Production Operations. In Machine Learning Applications in Subsurface Energy Resource Management. CRC Press: Boca Raton, FL, USA, 2022; pp. 313–336.

[13] Janssens, O.; Van de Walle, R.; Loccufier, M.; Van Hoecke, S. Deep learning for infrared thermal image based machine health monitoring. IEEE/ASME Trans. Mechatron. 2017, 23, 151–159.

[14] Scalabrini Sampaio, G.; Vallim Filho, A.R.d.A.; Santos da Silva, L.; Augusto da Silva, L. Prediction of motor failure time using an artificial neural network. Sensors 2019, 19, 4342.

[15] Bekar, E.T.; Nyqvist, P.; Skoogh, A. An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. Adv. Mech. Eng. 2020, 12, 1687814020919207.

[16] Otchere, D.A.; Ganat, T.O.A.; Gholami, R.; Lawal, M. A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. J. Nat. Gas Sci. Eng. 2021, 91, 103962.

[17] Susto, G.A.; McLoone, S.; Pagano, D.; Schirru, A.; Pampuri, S.; Beghi, A. Prediction of integral type failures in semiconductor manufacturing through classification methods. In Proceedings of the 2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA), Cagliari, Italy, 10–13 September 2013; pp. 1–4.

[18] Praveenkumar, T.; Saimurugan, M.; Krishnakumar, P.; Ramachandran, K.I. Fault diagnosis of automobile gearbox based on machine learning techniques. Procedia Eng. 2014, 97, 2092–2098.

[19] Kumar, A.; Shankar, R.; Thakur, L.S. A big data driven sustainable manufacturing framework for condition-based maintenance prediction. J. Comput. Sci. 2018, 27, 428–439. https://doi.org/10.1016/j.jocs.2017.06.006.

[20] Kolokas, N.; Vafeiadis, T.; Ioannidis, D.; Tzovaras, D. Forecasting faults of industrial equipment using machine learning classifiers. In Proceedings of the 2018 Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, Greece, 3–5 July 2018; pp. 1–6.

[21] Liu, L.; Liu, W. Calculation method of water injection forward modeling and inversion process in oilfield water injection network. AIP Conf. Proc. 2018, 1955, 040065.

[22] Aziz, N.; Abdullah, M.H.A.; Osman, N.A.; Musa, M.N.; Akhir, E.A.P. Predictive Analytics for Oil and Gas Asset Maintenance Using XGBoost Algorithm. In The International Conference on Emerging Technologies and Intelligent Systems; Springer International Publishing: Cham, Switzerland, 2022; pp. 108–117.

[23] Shehadeh, A.; Alshboul, O.; Al Mamlook, R.E.; Hamedat, O. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Autom. Constr. 2021, 129, 103827.

[24] Udo, W.; Muhammad, Y. Data-driven predictive maintenance of wind turbine based on SCADA data. IEEE Access 2021, 9, 162370–162388.

[25] Salim, K.; Hebri, R.S.A.; Besma, S. Classification predictive maintenance using XGboost with genetic algorithm. Rev. Intell. Artif. 2022, 36, 833.

[26] Hu, Z.; Chan, C.W. In-situ bioremediation for petroleum contamination: A fuzzy rule-based model predictive control system. Eng. Appl. Artif. Intell. 2015, 38, 70–78.

[27] Hosseinzadeh, A.; Chen, F.F.; Shahin, M.; Bouzary, H. A predictive maintenance approach in manufacturing systems via AI-based early failure detection. Manuf. Lett. 2023, 35, 1179–1186.

[28] Wang, M.; Shen, K.; Tai, C.; Zhang, Q.; Yang, Z.; Guo, C. Research on fault diagnosis system for belt conveyor based on internet of things and the LightGBM model. PLoS ONE 2023, 18, e0277352.

[29] Kizito, R.; Scruggs, P.; Li, X.; Kress, R.; Devinney, M.; Berg, T. The Application of Random Forest to Predictive Maintenance. In Proceedings of the 2018 IIE Annual Conference, Orlando, FL, USA, 19–22 May 2018; pp. 354–359; Institute of Industrial and Systems Engineers (IISE): Peachtree Corners, GA, USA, 2018.

[30] Chazhoor, A.; Mounika, Y.; Sarobin, M.V.R.; Sanjana, M.V.; Yasashvini, R. Predictive maintenance using machine learning based classification models. IOP Conf. Ser. Mater. Sci. Eng. 2020, 954, 012001.