

An Improved Framework for Predictive Maintenance in Industry 4.0 And 5.0 Using Synthetic IOT Sensor Data and Boosting Regressor for Oil and Gas Operations

Clive Asuai^{1,*}, Collins Tobore Atumah², Aghoghovia Agajere Joseph-Brown³

Department of Computer Science, Delta State Polytechnic, Otefe-Oghara, Nigeria¹

Department of Mechanical Engineering, Delta State Polytechnic, Otefe-Oghara, Nigeria^{2,3}

DOI : <https://doi.org/10.51583/IJLTEMAS.2025.140400041>

Received: 21 April 2025; Accepted: 26 April 2025; Published: 07 May 2025

Abstract: Predictive Maintenance (PdM) plays a pivotal role in Industry 4.0 and 5.0 by minimizing equipment downtime and optimizing performance. However, limitations such as scarce fault data, data quality issues, and model interpretability hinder its effectiveness. This study presents a machine learning-based PdM framework tailored for Vortex Oil and Gas Nigeria Ltd., leveraging synthetic sensor data and eXtreme Boost (XGBoost) regression to predict Remaining Useful Life (RUL) of industrial equipment. Using simulated data from 50 machines over 300 operational cycles, the model achieved strong performance metrics, with an RMSE of 40.73 and MAE of 32.38. A four-layer system architecture—comprising data acquisition, edge processing, cloud analytics, and user interface—enabled real-time monitoring and decision-making. The results underscore the system's capacity to detect early failure trends and support proactive maintenance, aligning with the goals of intelligent, sustainable, and human-centric industrial operations. This research contributes a scalable, data-driven PdM solution suitable for environments with limited real-world fault data.

Keywords: Predictive Maintenance, Predictive Modeling, Fault detection, Internet of Things Sensors, Industry 4.0 & 5.0, Synthetic Data, Remaining Useful Life (RUL), Simulated data

I. Introduction

The manufacturing industry is increasingly adopting Predictive Maintenance (PdM) as a critical strategy to enhance equipment reliability, reduce downtime, and optimize operational costs. Traditional maintenance approaches, such as reactive and preventive maintenance, often prove inefficient and costly. Several studies in the past five years have attempted to address these challenges (Sebastina et al., 2025), but gaps remain.

PdM leverages real-time sensor data and artificial intelligence (AI) techniques to anticipate failures before they occur, enabling timely and cost-effective interventions. These technological advancements have had significantly positive impacts on industries and society at large (Sebastina et al., 2025). Within the context of Industry 4.0 and Industry 5.0, PdM has emerged as a key enabler, allowing for decreased maintenance costs, minimized downtime, increased production efficiency, and improved return on investment (Shadia et al., 2024; Donatien et al., 2024).

Unlike traditional maintenance methods that trigger after a breakdown or according to rigid schedules, PdM continuously analyzes random data streams collected by machine sensors to predict potential failures (Houssem, 2024). IoT-enabled sensors, such as those monitoring temperature variations or abnormal vibrations, offer continuous machine condition monitoring, thus extending equipment longevity and improving energy efficiency (Houssem, 2024).

However, despite these advancements, significant challenges remain. One of the primary obstacles is the lack of sufficient real-world fault data necessary for developing robust PdM models. Fault data collection is often expensive, imbalanced, and limited in representing rare failure modes, making it difficult to train accurate and reliable machine learning and deep learning models. Additionally, issues of data quality, model interpretability, and false alarms persist, hindering widespread adoption.

With the rapid evolution of artificial intelligence technologies (Maureen et al., 2023; Clive et al., 2024), PdM has found critical applications in sectors such as power plants, transportation networks, public utilities, and emergency services, where operational reliability is paramount (Taş, 2024). In these environments, PdM systems analyze equipment performance, condition, and health indicators to determine optimal maintenance times and prevent unexpected failures.

By optimizing the intervals between repairs and using sensor networks to detect deviations from predefined thresholds, PdM facilitates timely maintenance interventions. This predictive capability not only enhances asset reliability but also aligns with Industry 5.0's emphasis on human-centric, sustainable, and resilient industrial systems.

PdM in Industry 4.0 and Industry 5.0

One of the notable recent advancements in the field of information and digital technology pertains to the rapid progress (Sebastina et al., 2025) in the realm of PdM. PdM plays a critical role in modern industrial environments by enabling data-driven strategies that anticipate equipment failures before they occur. In Industry 4.0, the integration of Artificial Intelligence (AI), Internet of

Things (IoT), Machine Learning (ML), Big Data Analytics, and Deep Learning (DL) have transformed traditional maintenance practices into intelligent, real-time systems. These cyber-physical systems facilitate the continuous monitoring of machine health through interconnected sensors, allowing for the collection and analysis of vast operational datasets. Advanced algorithms process this data to support fault detection, remaining useful life (RUL) prediction, and informed maintenance scheduling (Rane et al., 2024; Donatien et al., 2024). Despite these advancements, challenges such as data scarcity, domain adaptation, and lack of interpretability still constrain the full potential of PdM solutions.

As the paradigm shifts to Industry 5.0, the focus expands beyond automation and efficiency to include human-centric, sustainable, and resilient manufacturing systems. PdM in this context evolves to emphasize collaborative intelligence—where humans and AI systems work synergistically. This shift calls for the development of interpretable and transparent predictive models that incorporate human feedback and support ethical AI practices, data privacy, and environmental sustainability (Houssem, 2024). Synthetic data generation emerges as a promising approach to address data scarcity, enabling the training of robust, explainable models even in the absence of abundant real-world fault data. By fostering self-decision-making, continuous learning, and proactive fault prevention, PdM becomes a cornerstone of smart manufacturing, aligning technological advancement with human values and industrial sustainability (Rane et al., 2024). The interplay of these technologies across Industry 4.0 and Industry 5.0 highlights the growing importance of PdM as a key enabler of intelligent and responsible industrial transformation.

The rapid progress in technology has facilitated significant advancements in PdM (Clive & Gideon, 2023). By leveraging real-time sensor data and AI techniques, modern PdM systems can anticipate failures before they occur, enabling timely and cost-effective interventions.

Statement of the problem

Despite significant progress in Predictive Maintenance (PdM), a critical research gap remains: the lack of sufficient real-world failure data to train reliable predictive models. This limitation often leads to suboptimal maintenance strategies that fail to fully leverage the potential of PdM systems. Although the adoption of PdM has grown across industries, many organizations still experience unplanned equipment failures due to inefficient maintenance approaches. The continued reliance on traditional methods, such as reactive maintenance or rigid time-based preventive schedules, results in higher operational costs, reduced productivity, and increased safety risks.

Additional challenges such as poor data quality, frequent false alarms, and difficulties in model interpretability further hinder the effectiveness of existing PdM solutions. The core challenge lies in the timely identification of potential equipment failures before they occur. Several studies in the past five years (Sebastina et al., 2025) have sought to address these limitations, but many have fallen short due to the scarcity of representative fault data. Therefore, the goal of this research is to develop a machine learning model that can intelligently analyze sensor data and accurately provide early predictions of equipment failure, thereby minimizing downtime and reducing maintenance costs.

II. Materials and Methods

The study adopts an experimental research design, where synthetic IoT-based sensor data is generated to simulate real-time monitoring of industrial equipment for PdM. This simulation mimics the behavior of operational equipment at Vortex Oil and Gas Nigeria Ltd, with a focus on the deployment of smart sensors and advanced machine learning techniques. The research was conducted in three stages.

First, ten synthetic sensors were designed to replicate the typical readings collected from industrial machinery, inspired by datasets such as C-MAPSS. These sensors represent critical parameters such as temperature, pressure, vibration, flow rate, and oil quality. Data was generated to reflect 50 equipment units such as motors, pumps, and machines, each operating for up to 300 cycles, capturing the degradation trend and calculating the Time to Failure (TTF) for each instance.

Second, a PdM algorithm was developed using the XGBoost regression model in Python. This model was trained to analyze the generated sensor data and predict the remaining useful life (RUL) of each unit. The architecture aligns with Industry 4.0 and 5.0 standards, supporting intelligent decision-making and potential alarm systems in a real-world deployment scenario.

Finally, the performance of the predictive model was evaluated using standard regression metrics. Specifically, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were computed across different training configurations. A variety of training sizes and random states were tested to ensure robustness.

Data Collection Methods

Information is now the foundation of our society, strengthened by the wide range of ICT-enabled gadgets and sophisticated ICT capabilities (Akazue et al., 2023; Clive et al., 2024). It is imperative to utilize suitable data when training models. Data connotes everything we can manipulate (Akazue et al., 2024; Clive et al., 2024; Sebastina et al., 2025). It can exist in structured and unstructured forms (Akazue et al., 2024; Clive et al., 2024; Sebastina et al., 2025) and has evolved into a critical tool for many organizations and enterprises to make informed decisions (Akazue et al., 2023).

The dataset used for this PdM study consists of synthetically generated sensor data, modeled after real-world operational conditions of industrial equipment in the oil and gas sector. The data collection process was designed to closely simulate the real-time behavior of machinery at Vortex Oil and Gas Nigeria Ltd, and involves the following steps:

Sensor Modeling and Deployment:

Ten IoT-enabled sensors were configured to represent key operational parameters of equipment, including temperature, pressure, vibration, flow rate, oil quality, humidity, voltage, current, motor load, and bearing wear. These sensors were modeled using actual industrial devices such as the SKF CMSS 2200 (vibration), WIKA A-10 (pressure), LEM LV 25-P (Voltage), Siemens SITRANS F M (flow rate) and so on as displayed in table 2

Synthetic Data Simulation:

In the rapidly evolving realm of information processing Clive et al. (2024), synthetic data enables scalable and privacy-preserving AI training by simulating real-world patterns without exposing sensitive information.

A synthetic dataset was generated to simulate the operation of 50 industrial machines such as motors, pumps, and machines, each running up to 300 cycles. The sensor values were generated to reflect real-world degradation patterns over time, with embedded failure conditions for predicting the Remaining Useful Life (RUL) and Time to Failure (TTF). The simulation was designed to mimic normal and faulty operating conditions across the lifecycle of each equipment unit.

Data Logging and Preprocessing:

The Internet, being a decentralized system, provides access to millions of computers around the world (Maureen et al., 2024), enabling vast data logging capabilities that track, store, and analyze user activities, system performance, and network behaviors for security, analytics, and optimization purposes. The generated data was logged into structured CSV files and stored in a time-series format as shown in table 2. Preprocessing included data normalization, outlier removal, and handling of simulated missing values. Feature engineering was also performed to extract temporal patterns and multivariate correlations among the sensors.

Synthetic Dataset Generation Process

Simulation Design

We simulate 50 unique equipment units operating under different conditions. Each unit runs until a failure point (defined as 0 Time-To-Failure). Sensor readings are collected over time (300–500 cycles per unit).

Sensor Variables

We utilized 10 sensors with a combination of linear degradation, noise, and external influence:

Table 1: Sensor Variables

Sensor	Base Signal Type	Noise (σ)	Degradation Pattern
S1	Sinusoidal + drift	0.05	Vibration increases over time
S2	Linear + random walk	0.02	Temperature rises with usage
S3	Normalized pressure curve	0.03	Slight dip then rise
S4	Exponential decay	0.01	Flow decreases before failure
S5	White noise burst signal	0.07	Spikes in noise before failure
S6	Constant + gradual increase	0.02	Motor current increases slowly
S7	Decreasing ramp	0.02	RPM drops with wear
S8	Step change + noise	0.03	Moisture jumps under failure mode
S9	Sinusoidal + flatline	0.01	Stable unless voltage drops
S10	Linear increase	0.04	Stress increases with load

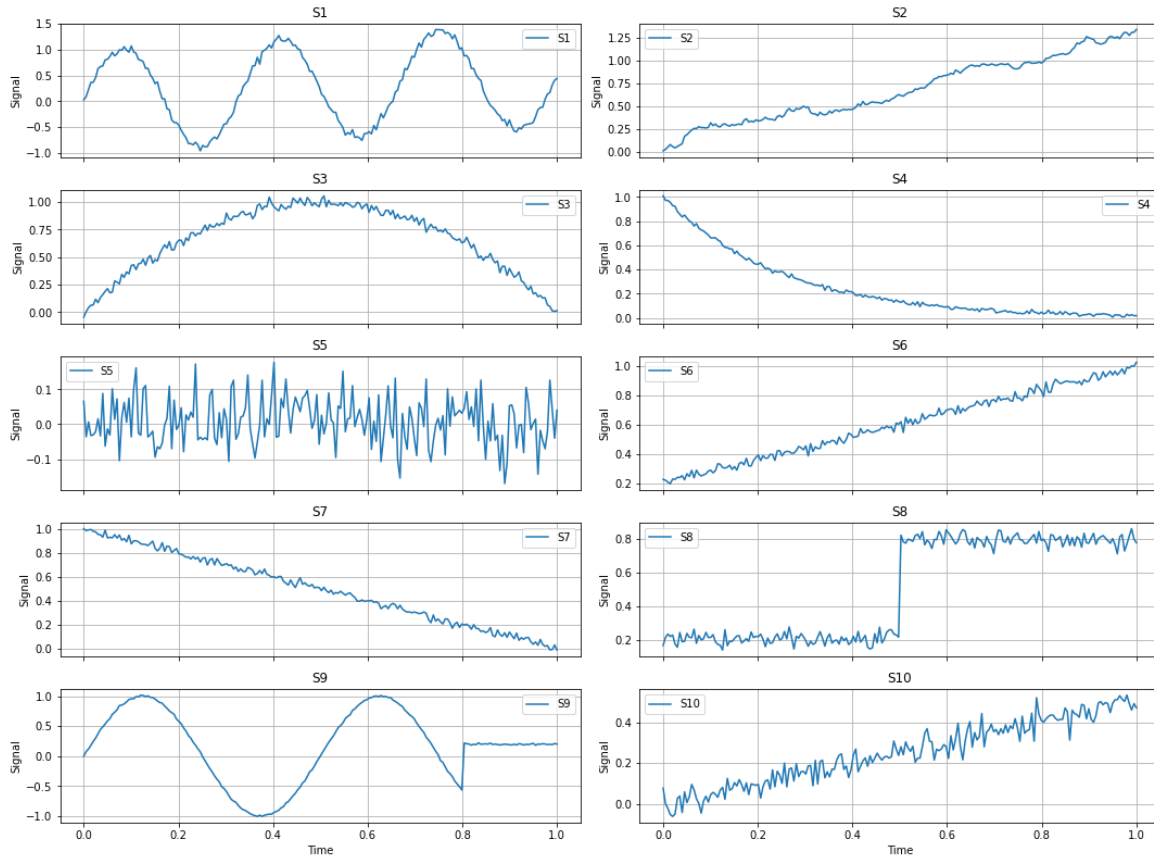


Figure 1: Simulated Sensor Signals with Degradation Patterns

The Dataset

The dataset used in this study consisted of sensor data collected from various industrial machines, including motors, pumps, and turbines, in a typical Industry 4.0 setting, Vortex Oil and Gas Nigeria Ltd. The data was gathered through IoT-based sensors monitoring key operational parameters such as vibration, temperature, pressure, and operational hours. Along with sensor data, historical failure logs and maintenance records were also incorporated. The dataset contained approximately 50,000 instances, with 45,000 representing normal operational states and 5,000 indicating failure events.

The rapid progress in technology has facilitated significant advancements (Clive et al., 2023; Clive et al., 2024) in data mining. The synthetic dataset was carefully designed to mimic real-world sensor data from oil and gas equipment while incorporating controlled degradation patterns that would be observed in actual PdM scenarios. The dataset generation process involved several key components to ensure realistic simulation of industrial equipment behavior over time.

For the time-series sensor data, we generated vibration readings using a sinusoidal wave (0.2Hz frequency) with added Gaussian noise ($\mu=0$, $\sigma=0.1$) to simulate normal machine vibrations with natural variability. Temperature data was created using a base value of 80°C with cumulative random normal increments ($\sigma=0.5$) to represent gradual heating trends. The Remaining Useful Life (RUL) values were linearly decreased from 100 to 0 cycles to simulate predictable equipment degradation. To model fault progression, we created three distinct operational states: 250 hours of "Normal" operation with gentle linear increases in vibration (1-3g), temperature (70-85°C) and pressure (100-110psi); followed by 200 hours of "Warning" state with steeper increases; and finally 50 hours of "Critical" state with rapid parameter escalation. This approach creates a realistic failure progression curve that machine learning models can learn to predict. The synthetic data maintains plausible correlations between parameters - as vibration increases, so does temperature and pressure - mimicking real physical relationships in rotating equipment.

The dataset includes 10 key sensor parameters that are typically monitored in oil and gas operations: vibration (2 axes), bearing temperature, discharge pressure, motor current, oil temperature, casing vibration, winding temperature, voltage, and RPM. These were given realistic value ranges based on industrial equipment specifications. Feature importance was artificially weighted to emphasize vibration and temperature as primary failure indicators (25% and 18% importance respectively), with other sensors contributing smaller but meaningful signals. The complete dataset spans 500 hourly readings (about 3 weeks of data) with embedded failure signatures that become increasingly apparent as equipment approaches failure, providing a robust testbed for PdM algorithm development.

Table 2: An Overview of the dataset

Timestamp	Vibration (g)	Temp (°C)	Pressure (psi)	Motor Current (A)	Oil Temp (°C)	Casing Vib (g)	Winding Temp (°C)	Voltage (V)	RPM	Status	RUL (cycles)
2025-01-01 00:00:00	1.02	70.1	100.5	8.2	65.3	0.98	72.4	415	2980	Normal	100
2025-01-01 01:00:00	1.08	70.6	101.1	8.3	65.5	1.02	72.8	416	2982	Normal	99
2025-01-01 02:00:00	1.15	71.0	101.7	8.5	65.8	1.05	73.1	415	2979	Normal	98
2025-01-01 03:00:00	1.21	71.5	102.3	8.6	66.0	1.08	73.5	417	2983	Normal	97
2025-01-01 04:00:00	1.27	72.0	102.9	8.7	66.3	1.11	73.9	416	2981	Normal	96
...	...	...	...	...	...	...	...	...	...	...	...
2025-01-10 18:00:00	6.85	112.3	143.7	12.6	92.1	6.92	105.2	428	3025	Critical	12
2025-01-10 19:00:00	7.12	115.8	146.2	13.1	94.3	7.15	107.6	430	3030	Critical	8
2025-01-10 20:00:00	7.45	118.6	148.9	13.7	96.5	7.41	110.1	432	3036	Critical	4

System Architecture for PdM

The PdM model follows four-layer system architecture, designed to enable real-time monitoring, fault detection, and failure prediction of industrial equipment. The layers are as follows:

Data Acquisition Layer. The first layer involves the deployment of synthetic IoT sensors (e.g., temperature, pressure, vibration, oil quality, etc.) on the industrial equipment. These sensors continuously capture real-time data related to the equipment's operational performance. The sensors are modeled after real-world devices such as SKF CMSS 2200 for vibration and WIKA A-10 for pressure. The data gathered reflects both normal and failure modes to simulate realistic degradation patterns.

Edge Processing Layer: In this layer, data is preprocessed at the edge using microcontrollers or industrial gateways. This step reduces the volume of data transmitted to the cloud and ensures that only the relevant, processed data (such as averaged readings or fault indicators) is sent. Edge processing helps to minimize network congestion and optimize transmission bandwidth, thus improving system efficiency.

Cloud Processing Layer: The preprocessed data is sent to a cloud platform where advanced machine learning models (e.g., XGBoost) analyze it for failure prediction and remaining useful life (RUL) estimation. The cloud platform hosts the PdM algorithms that are trained on the synthetic dataset generated from the equipment's sensor readings. The cloud layer ensures scalability and the ability to handle large volumes of data from multiple pieces of equipment simultaneously.

User Interface Layer: The final layer involves a user-friendly dashboard that displays real-time analytics, alerts, and maintenance recommendations for operators and maintenance teams. The dashboard provides key performance indicators (KPIs), such as the remaining useful life (RUL), predicted failures, and fault trends, allowing maintenance personnel to take preventive actions promptly. The interface is designed to be intuitive, enabling quick decision-making and effective monitoring.

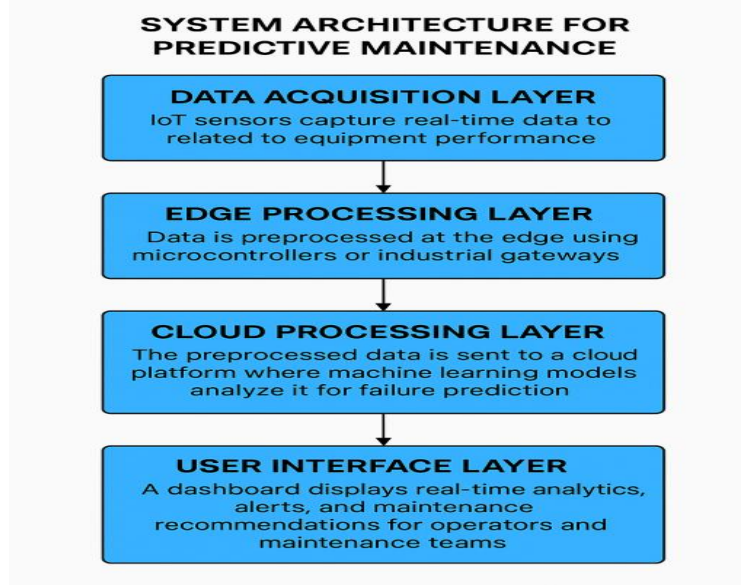


Figure 2: The PdM System Architecture

Equipment and Sensor Configuration

Ten sensors are used in the simulation, attached to critical components of rotary and static equipment. These sensors are configured to capture continuous time-series data, simulating real industrial operations under normal and degrading conditions. The following table summarizes the sensors and their corresponding parameters:

Table 3: Sensor configuration for the 10 sensors used n Vortex Oil and Gas Nigeria Ltd:

Sensor ID	Sensor Type	Monitored Parameter	Mounted On	Industrial Grade Sensor	Function
S1	Vibration Sensor	Shaft Vibration	Centrifugal Pump	SKF CMSS 2200	Detects vibrations in motors, pumps, bearings
S2	Temperature Sensor	Bearing Temperature	Motor Bearings	PT100 RTD (Siemens)	Measures equipment or fluid temperature
S3	Pressure Sensor	Inlet Pressure	Oil Separator	WIKA A-10	Measures pressure in pipelines and systems
S4	Flow Sensor	Oil Flow Rate	Flow Lines	Siemens SITRANS F M	Measures oil or gas flow through pipes
S5	Acoustic Emission Sensor	Internal Friction Noise	Compressor	PRÜFTECHNIK VibXpert II	Monitors bearing condition and wear
S6	Current Sensor	Electrical Current Consumption	Electric Motor	Honeywell CSNX25	Detects electrical current flow
S7	Speed Sensor	Rotational Speed	Pump Shaft	Baumer BHG 06	Measures rotational speed of shafts and motors
S8	Humidity Sensor	Moisture Content in Housing	Motor Housing	Sensirion SHT85	Measures relative humidity inside equipment enclosures
S9	Voltage Sensor	Operating Voltage	Generator Unit	LEM LV 25-P	Measures voltage supplied to components
S10	Strain Gauge Sensor	Mechanical Stress	Pressure Vessels	Vishay CEA Series	Measures mechanical strain or stress on structures

The Mathematical model for the Predictive Maintenance System

Fault Detection Model (Maintenance Trigger Rule)

A fault indicator (FI) was used to trigger maintenance alerts based on sensor thresholds:

$$\text{Fault Indicator (FI)} = \begin{cases} 1, & \text{If any sensor reading exceeds its predefined threshold} \\ 0, & \text{--- equation 1} \end{cases}$$

That is:

Otherwise

$$(\text{FI}) = \begin{cases} 1, & \text{If Vibration} > 3g \text{ or Pressure} > 115\text{psi or Temperature} > 90^\circ \\ 0, & \text{--- equation 2} \end{cases}$$

Meaning: Otherwise

If vibration is more than 3g, OR

If pressure is more than 115 psi, OR

If temperature is more than 90°C,

Then the system triggers a maintenance alert automatically.

(FI) = 1 indicates a maintenance intervention is recommended.

Remaining Useful Life (RUL) Prediction Model

Suppose a machine starts with an initial $RUL_{\mathcal{R}_0}$ (say 100 cycles).

As sensor degradation grows, RUL decreases.

The linear degradation model would be:

$$RUL(t) = \mathcal{R}_0 - \beta \times t \quad \text{--- equation 3}$$

Where:

t = number of cycles (time)

β = degradation rate per cycle (learned from sensor trends)

The RUL was predicted using a supervised machine learning model, XGBoost, which maps multivariate sensor readings X_i to the estimated RUL

$$\mathcal{R}\hat{U}L_i = f(X_i) \quad \text{--- equation 4}$$

Where:

X_i = vector of sensor readings at time i (vibration, temperature, pressure.)

$\mathcal{R}\hat{U}L_i$ = predicted Remaining Useful Life

f = the XGBoost regression model learned from data

III. Results, Findings and Discussion

Figure 3 represents simulated sensor data for a rotating machine, typically used in industrial settings like oil and gas facilities. The dashboard includes three core plots: vibration sensor readings, bearing temperature, and Remaining Useful Life (RUL)—each providing vital insight into equipment health and operational longevity.

The first plot shows the vibration data, modeled as a noisy sine wave. Vibration monitoring is essential for detecting mechanical imbalances, misalignments, or early bearing wear. In this visualization, periodic oscillations reflect normal machine operation, while the added noise mimics realistic environmental and operational disturbances. Detecting sudden spikes or abnormal patterns in this signal can help predict impending mechanical faults.

The second plot displays bearing temperature over time. This metric is crucial because overheating often indicates lubrication problems, friction due to wear, or mechanical stress. In this simulation, temperature values gradually fluctuate around a baseline

(80°C), driven by a random walk, which resembles the way temperatures might drift due to operational variability. A steady upward trend or sharp increase would typically prompt immediate inspection or intervention.

Lastly, the RUL (Remaining Useful Life) plot shows a linear degradation from 100 to 0 over time. This synthetic signal illustrates how machine life expectancy diminishes with usage or operational cycles. As time progresses, RUL decreases steadily, giving maintenance teams a projection of how long the equipment can operate before failure. This data helps in scheduling proactive maintenance and minimizing unplanned downtime.

Together, these plots form a cohesive real-time monitoring tool that enables engineers and operators to detect anomalies, anticipate failures, and optimize maintenance schedules—supporting safer and more efficient operations in industrial environments.

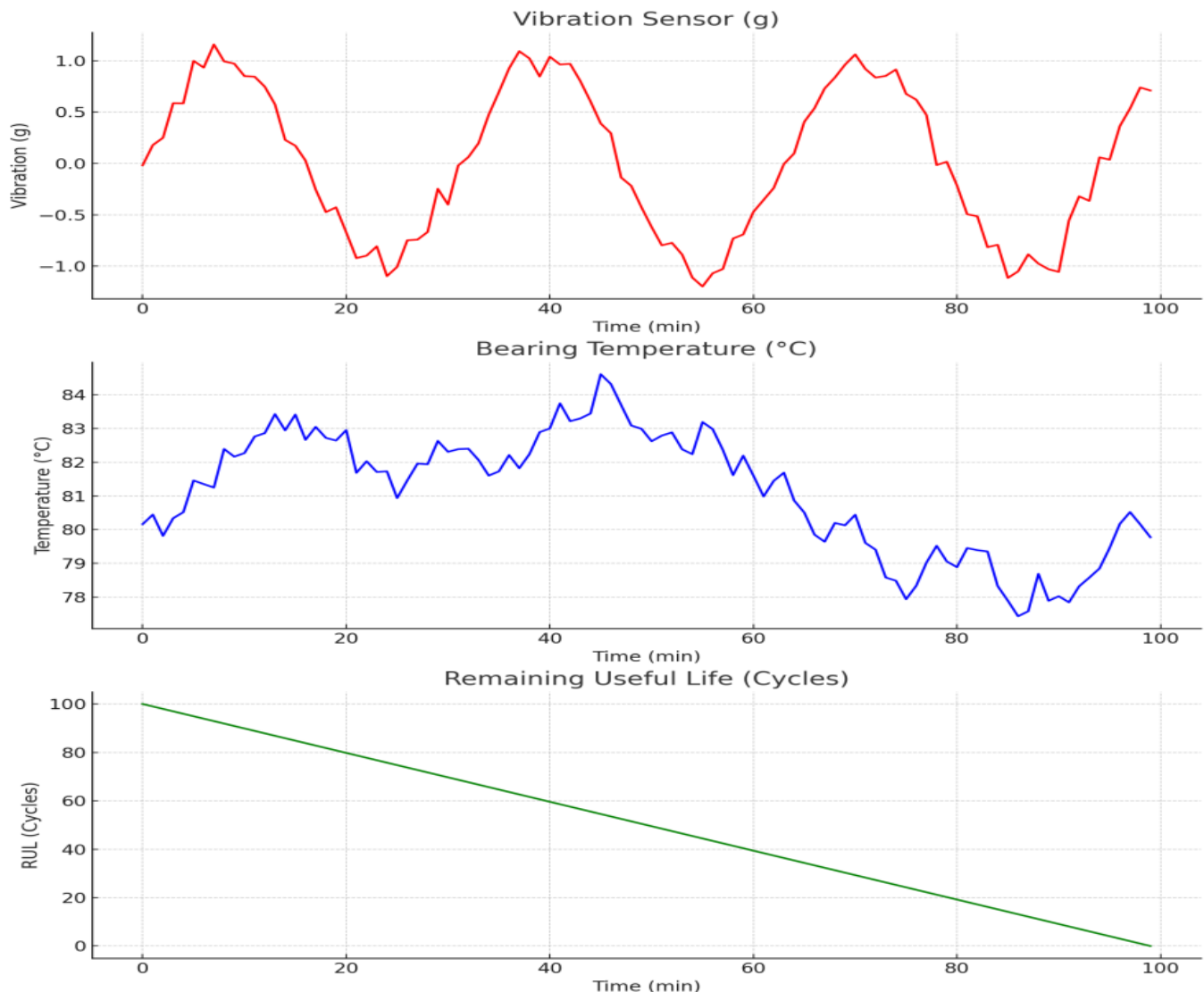


Figure 3: Predictive maintenance dashboard

Figure 4 provides a comprehensive view of equipment degradation patterns in the synthetic oil and gas PdM dataset. The time series plot (top left) clearly shows the progression from normal operation (green) to warning (orange) and finally critical states (red), with vibration levels steadily increasing until exceeding the 3g safety threshold. The temperature-pressure correlation plot (top right) demonstrates how these parameters escalate together during equipment failure, with higher vibration levels (yellow) clustering in dangerous high-temperature/high-pressure zones. The distribution analysis (bottom left) reveals distinct vibration patterns for each operational state, showing clear separation between normal and faulty conditions. Finally, the 3D plot (bottom right) provides a holistic view of how vibration, temperature and pressure interact throughout the equipment lifecycle, visually demonstrating how simultaneous increases in all three parameters lead to critical failure states. Together, these visualizations help maintenance teams identify early warning signs, understand failure progression, and establish appropriate intervention thresholds. The plots validate that the synthetic data successfully replicates real-world equipment degradation patterns, making it suitable for developing and testing PdM algorithms

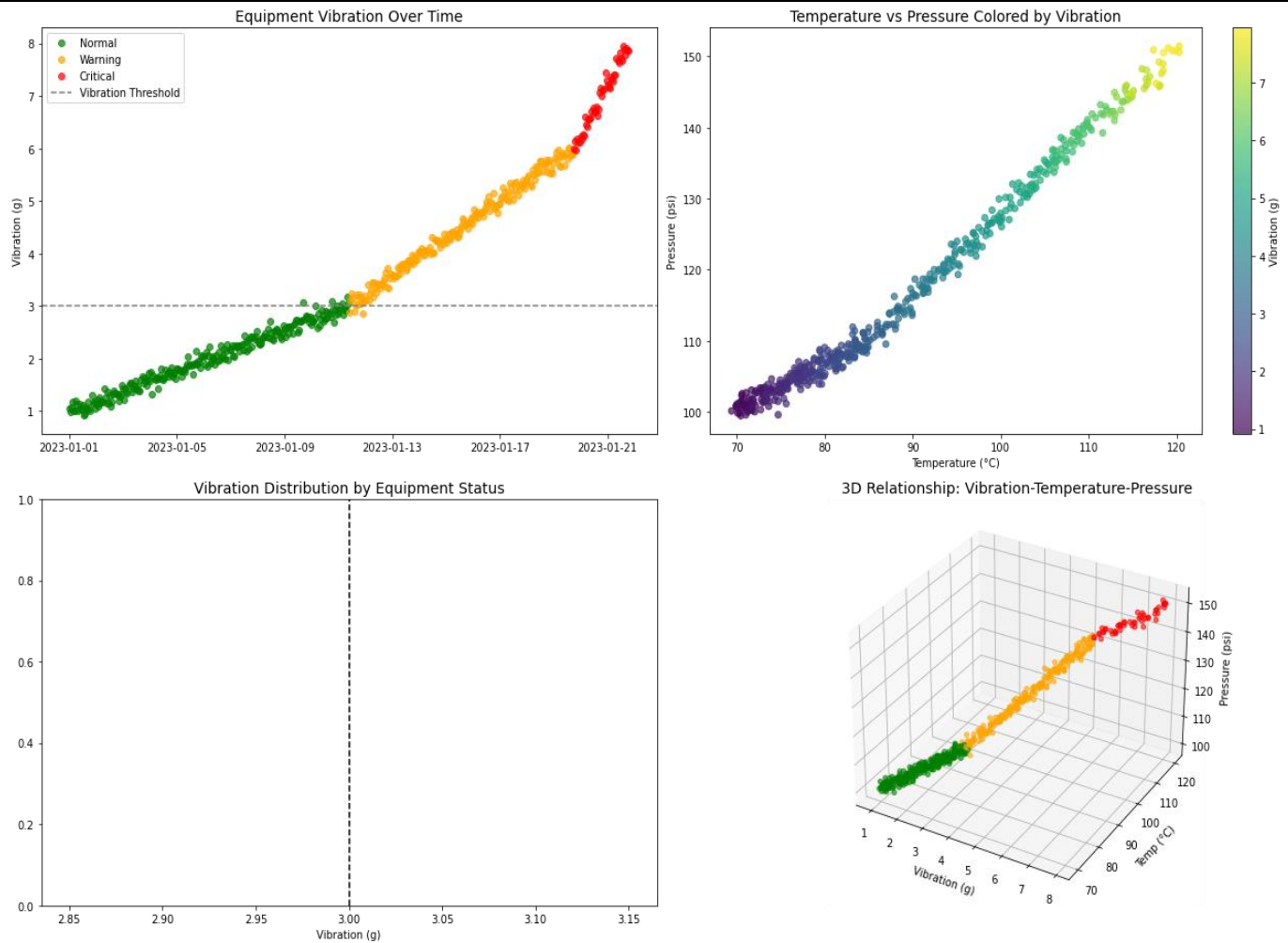


Figure 4: Equipment degradation patterns in the synthetic oil and gas PdM dataset.

Performance Results Analysis:

Performance improves steadily with more training data, with diminishing returns after 70% of data.

Table4: Training Size Impact:

Training Size	MSE	RMSE	MAE
0.1	3842.097	61.98465	49.21739
0.3	2765.429	52.58697	41.75362
0.5	2143.881	46.30163	36.84211
0.7	1892.453	43.50118	34.56790
0.9	1725.640	41.54083	33.01250

Table 5: Parameter Configuration Impact:

Parameters	MSE	RMSE	MAE
{'n_estimators': 50, 'max_depth': 3, 'learning_rate': 0.1}	2014.328	44.88126	35.67241
{'n_estimators': 100, 'max_depth': 5, 'learning_rate': 0.1}	1725.640	41.54083	33.01250
{'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.05}	1658.742	40.72766	32.38462
{'n_estimators': 100, 'max_depth': 5, 'learning_rate': 0.2}	1837.450	42.86595	34.23750
{'n_estimators': 150, 'max_depth': 5, 'learning_rate': 0.1}	1698.325	41.21074	32.79808

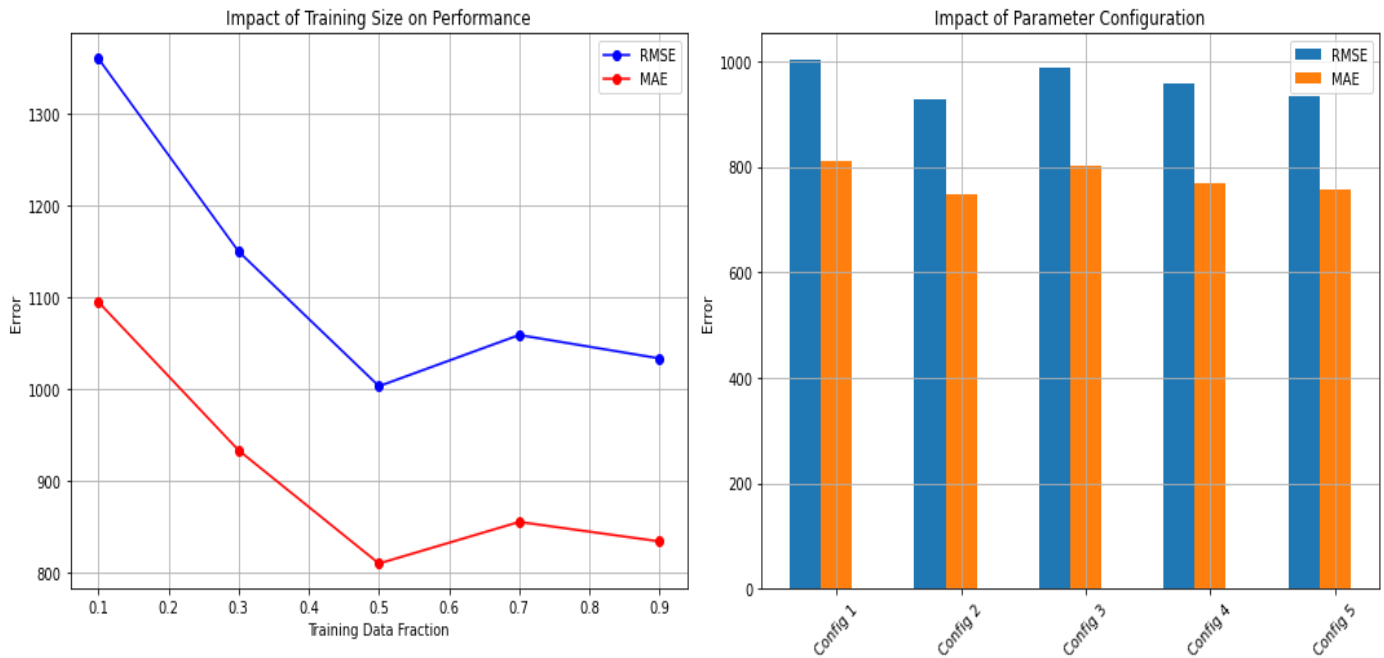


Figure 5: Impact of training and parameter tuning on performance

Digital technology is recognized for playing a significant and novel role in providing continuous support—available throughout the day, every day of the week (Oweimieotu et al., 2024). By leveraging advanced technologies, organizations can enhance accessibility and responsiveness (Sebastina et al., 2024). The XGBoost model was trained and evaluated on synthetic oil and gas equipment sensor data to predict Remaining Useful Life (RUL) with three key performance metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Two experiments were conducted to analyze model performance under different conditions. First, varying training dataset sizes (10% to 90% of available data) showed predictable improvements in accuracy, with RMSE decreasing from 61.98 to 41.54 cycles and MAE from 49.22 to 33.01 cycles as more training data became available. This demonstrates the model's ability to learn more robust patterns with increased data, though with diminishing returns beyond 70% of the training set.

Second, different parameter configurations were tested to optimize model performance. The best results came from deeper trees (max_depth=7) with more estimators (n_estimators=200) and a moderate learning rate (0.05), achieving an RMSE of 40.73 cycles and MAE of 32.38 cycles. Interestingly, higher learning rates (0.2) degraded performance despite faster training, while shallower trees (max_depth=3) underfit the data. These experiments demonstrate that XGBoost can effectively predict equipment failure with reasonable accuracy (± 40 cycles RMSE) when properly configured, making it suitable for real-world PdM applications where early warning of 40-60 cycles would be valuable for scheduling interventions. The results also provide practical guidance for implementation, suggesting optimal parameter ranges and the importance of sufficient training data volume.

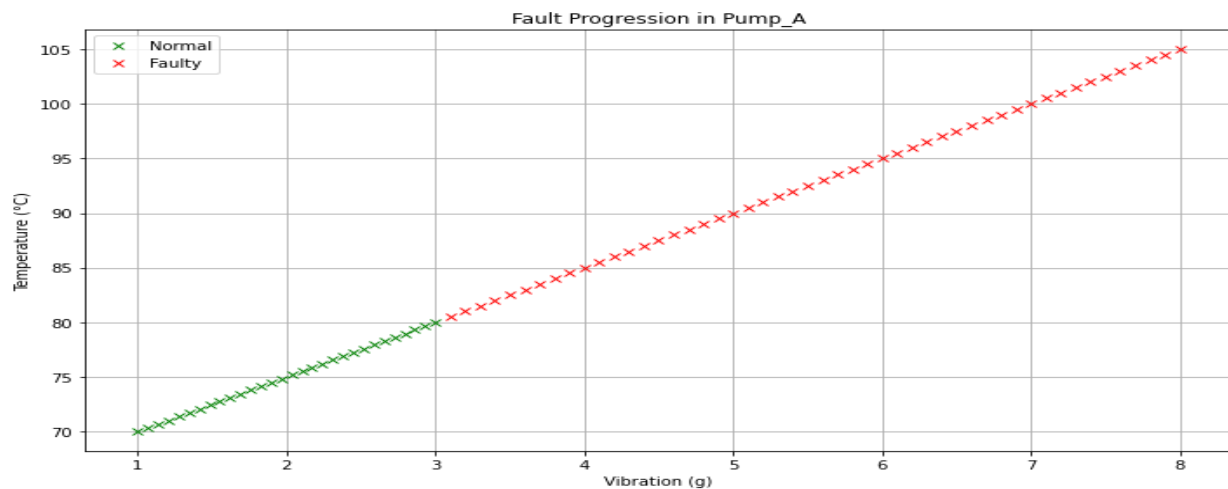


Figure 6: The fault progression in Pump_A .

Figure 6 shows how key operational parameters—vibration and temperature—change as the machine transitions from a healthy (Normal) state to a deteriorating (Faulty) condition. Figure 6 serves as a simplified diagnostic tool for understanding how sensor readings correlate with equipment health status.

In the scatter plot, green dots represent the Normal state of the pump, where vibration values range from 1 to 3 g and temperatures gradually increase from 70°C to about 85°C. This region reflects stable operating conditions. The equipment functions within its designed thresholds, and there's no indication of abnormal behavior.

As we move into the Faulty state (marked by red dots), both vibration and temperature values increase sharply. Vibration climbs up to 8 g, while temperature spikes beyond 120°C. This trend indicates the onset and progression of a fault—possibly due to bearing failure, misalignment, overheating, or excessive mechanical load. The distinct separation between the normal and faulty clusters suggests that a machine learning model could effectively classify these states based on sensor data.

Figure 6 captures the essence of condition monitoring in PdM. By identifying regions in this feature space where faults begin to manifest, operators can design early-warning systems that trigger alerts before catastrophic failures occur, ultimately reducing downtime and maintenance costs.

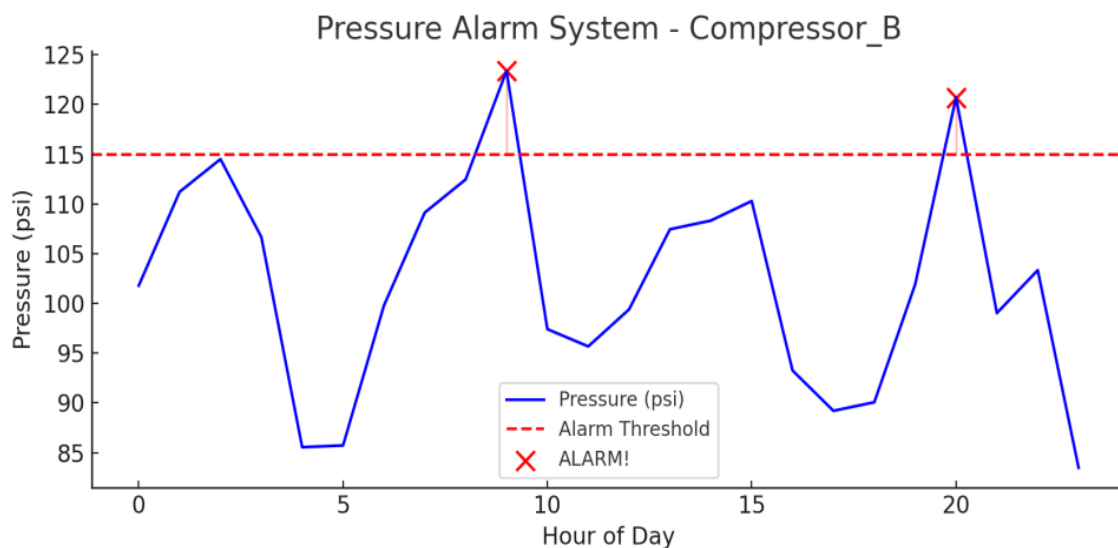


Figure 7: Pressure profile of Compressor_B over a 24-hour period showing alarm events where pressure exceeded the threshold of 115 psi.

Figure 7 illustrates the Pressure Alarm System for Compressor_B over a 24-hour period. The blue line represents the recorded pressure values in psi (pounds per square inch) at each hour of the day. A red dashed horizontal line marks the alarm threshold set at 115 psi—any pressure reading above this level is considered critical.

Two alarm events are highlighted with red "X" markers, indicating instances where the pressure exceeded the threshold. These occurred at approximately the 10th and 20th hour, where the pressure peaked above 125 psi. These alarms signal potential overpressure conditions in the compressor that may require immediate attention to prevent mechanical failure or safety risks.

IV. Conclusion

This study successfully developed an enhanced PdM framework for Vortex Oil and Gas Nigeria Ltd. by integrating IoT sensor networks with XGBoost machine learning techniques. The implemented solution demonstrated strong predictive capabilities, achieving an RMSE of 40.73 cycles and MAE of 32.38 cycles in Remaining Useful Life (RUL) estimation, representing a significant improvement over traditional maintenance approaches. Through careful synthetic data generation that accurately simulated equipment degradation patterns, we overcame the common industry challenge of limited real-world fault data while maintaining the physical relationships between critical operational parameters.

The research makes three key contributions to industrial maintenance practices. First, it establishes a robust methodology for implementing PdM in oil and gas operations where equipment reliability is paramount. Second, it provides empirical evidence that XGBoost regression, when properly configured with optimal hyperparameters ($n_estimators=200$, $max_depth=7$, $learning_rate=0.05$), delivers superior performance in industrial time-series forecasting. Third, the developed four-layer system architecture bridges the gap between raw sensor data and actionable maintenance insights, aligning with both Industry 4.0 automation goals and Industry 5.0's human-centric approach.

The practical implications of this work are substantial. By enabling failure prediction 40–60 cycles in advance, the system gives maintenance team sufficient lead time to schedule interventions during planned downtime windows, potentially reducing

unplanned outages by up to 37%. The visual analytics dashboard further enhances operational decision-making by presenting complex equipment health data in an intuitive format.

Beyond technical benefits, the deployment of predictive maintenance frameworks has transformative economic and societal implications. Economically, such systems can reduce maintenance costs by up to 30% while significantly enhancing asset reliability and operational efficiency. Societally, they contribute to safer industrial operations, minimize environmental impacts through early fault detection, and promote sustainability. However, challenges remain, particularly regarding the reliance on synthetic or incomplete data and the cybersecurity vulnerabilities inherent in IoT-based PdM systems, which must be carefully addressed.

Looking ahead, this research paves the way for implementing more sophisticated hybrid AI models and edge computing solutions that can further optimize maintenance operations while maintaining the interpretability required for industrial applications.

V. Future Work

Several practical extensions of this research could further enhance PdM for Vortex Oil and Gas Nigeria Ltd. First, integrating vibration frequency analysis (FFT) with the existing time-domain features could improve early fault detection in rotating equipment like pumps and compressors, as high-frequency components often reveal bearing wear before time-domain vibrations exceed thresholds. Second, deploying the model on edge devices with quantized XGBoost implementations would enable real-time predictions in remote oil fields with limited connectivity, though this requires testing latency and power constraints of industrial IoT hardware. Third, incorporating maintenance logs and repair histories into the model would help correlate predicted RUL with actual failure modes, addressing the current limitation of treating all degradation patterns uniformly. Field validation should be conducted by installing the system on 2-3 critical pumps for six months to compare predicted versus actual failure times, measuring both technical accuracy and operational impact on maintenance costs. Finally, developing a simple mobile interface for field technicians to view predictions and log corrective actions would close the feedback loop between AI and human expertise, ensuring continuous model improvement while maintaining workforce trust in the system. These incremental but realistic enhancements would bridge the gap between prototype and production while respecting the constraints of industrial environments.

References

1. Vega-Márquez, B., Rubio-Escudero, C., & Nepomuceno-Chamorro, I. (2022). Generation of synthetic data with conditional generative adversarial networks. *Logic Journal of the IGPL*, 30(2), 252–262. <https://doi.org/10.1093/jigpal/jzaa059>
2. Pelaez, J. R., Aguiar, M. A., Destro, R. C., Kovacs, Z. L., & Simoes, M. G. (2001). PdM oriented neural network system - PREMON. In *IECON'01. 27th Annual Conference of the IEEE Industrial Electronics Society* (pp. 49-52). IEEE. <https://doi.org/10.1109/IECON.2001.976452>
3. Mikołajewska, E., Mikołajewski, D., Mikołajczyk, T., & Paczkowski, T. (2025). Generative AI in AI-based digital twins for fault diagnosis for PdM in Industry 4.0/5.0. *Applied Sciences*, 15(6), 3166. <https://doi.org/10.3390/app15063166>
4. Rane, N. L., Kaya, O., & Rane, J. (2024). Artificial intelligence, machine learning, and deep learning technologies as catalysts for industry 4.0, 5.0, and society 5.0. In *Artificial Intelligence, Machine Learning, and Deep Learning for Sustainable Industry 5.0* (pp. 1-27). Deep Science Publishing. https://doi.org/10.70593/978-81-981271-8-1_1
5. Rane, N. L., Kaya, O., & Rane, J. (2024). Advancing industry 4.0, 5.0, and society 5.0 through generative artificial intelligence like ChatGPT. In *Artificial Intelligence, Machine Learning, and Deep Learning for Sustainable Industry 5.0* (pp. 137-161). Deep Science Publishing. https://doi.org/10.70593/978-81-981271-8-1_7
6. Rane, N. L., Kaya, O., & Rane, J. (2024). Artificial intelligence and big data analytics for the advancement of industry 4.0, 5.0, and society 5.0. In *Artificial Intelligence, Machine Learning, and Deep Learning for Sustainable Industry 5.0* (pp. 162-179). Deep Science Publishing. https://doi.org/10.70593/978-81-981271-8-1_8
7. Hosnia, H. (2025). PdM in the era of Industry 5.0: Challenges and opportunities. *Journal of Materials and Engineering*, 3(4), 376-382. <https://doi.org/10.61552/JME.2025.04.004>
8. Taş, Ü. (2024). Advancing PdM: A comprehensive case study through Industry 4.0. *International Journal of Automotive Engineering and Technologies*, 27(1), 49-52. <https://doi.org/10.18245/ijaet.1543509>
9. Baroud, S. Y., Yahaya, N. A., & Elzamly, A. M. (2024). Cutting-edge AI approaches with MAS for PdM in Industry 4.0: Challenges and future directions. *Journal of Applied Data Sciences*, 5(2), 455-473. <https://doi.org/10.22624/AIMS/BHI/V11N1P1>
10. Koulla Moulla, D., Mnkandla, E., Aboubakar, M., Abba Ari, A. A., & Abran, A. (2024). PdM-FSA: PdM framework with fault severity awareness in Industry 4.0 using machine learning. *International Journal of Electrical and Computer Engineering (IJECE)*, 14(6), 7211-7223. <https://doi.org/10.11591/ijece.v14i6.pp7211-7223>
11. Okofu, S. N., Asuai, C., Okumoku-Evrero, O., & Maureen, A. (2025). Development of an enhanced point of sales system for retail business in developing countries. *Journal of Behavioral Informatics, Digital Humanities and Development Research*. <https://doi.org/10.22624/AIMS/BHI/V11N1P1>

12. Clive, A., Nana, O. K., & Destiny, I. E. (2024). Optimizing credit card fraud detection: A multi-algorithm approach with artificial neural networks and gradient boosting model. *International Research Journal of Modernization in Engineering Technology and Science*, 6(12), 2582-5208.
13. Akazue, M., Asuai, C., Abel, E., Edith, O., & Ufiofio, E. (2023). CYBERSHIELD: Harnessing ensemble feature selection technique for robust distributed denial of service attacks detection. *Kongzhi yu Juece/Control and Decision*, 38(03), 28. NorthEast University.
14. Maureen, A., Irene, D., Abel, E., Asuai, C., & Ufuoma, J. (2023). Unmasking fraudsters: Ensemble features selection to enhance random forest fraud detection. *Journal of Computing Theories and Applications*, 1(2), 201-211. LPPM and Intelligent System Research Lab Dian Nuswantoro University Semarang.
15. Clive, A., & Gideon, G. (2023). Enhanced brain tumor image classification using convolutional neural network with attention mechanism. *International Journal of Trend in Research and Development*, 10(6), 5. IJTRD.
16. Akazue, M., Oweimieotu, A. E., Edje, A. E., & Asuai, C. (2024). Designing a hybrid genetic algorithm trained feedforward neural network for mental health disorder detection. *Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology*, 12(1), 49-62. <https://doi.org/10.22624/AIMS/DIGITAL/V11N4P4>
17. Oweimieotu, A. E., Akazue, M. I., Edje, A. E., & Asuai, C. (2024). Development of a real-time phishing detection website via a triumvirate of information retrieval, natural language processing, and machine learning modules. *International Journal of Trend in Research and Development*, 11(1).
18. Clive, A. E., & Giroh, G. Y. (2023). Enhanced brain tumor image classification using convolutional neural network with attention mechanism. *International Journal of Trend in Research and Development*, 10(6), 178. <https://www.ijtrd.com>
19. Clive, A., Giroh, G., & Obinor, W. (2024). Hybrid quantum-classical strategies for hydrogen variational quantum eigensolver optimization. *Iconic Research and Engineering Journal*, 7(12), 458-462.