Jurnal Personalia Pelajar 29(1): 83-94
https://doi.org/10.17576/personalia.2901.2026.09
ISSN 0128-2735

Ai-Powered Total Preventive Maintenance with LLM
(Penyelenggaraan Pencegahan Total Berkuasa AI dengan LLM)

SUTHAREESANAN SARAVANAN, SITI AZIRAH ASMAI, ZURAIDA ABAL ABAS, HAZLI RAFIS ABDUL RAHIM, SABRI MOHAMAD SHARIF, MOHD SYAFIQ ABD AZIZ & MOHD KHAIRUL NIZAM SUHAIMIN

ABSTRACT

This research-based project aims to mitigate unplanned downtime and improve maintenance efficiency in industrial settings, driven by the significant financial and operational penalties associated with unexpected equipment failure. Its novelty lies in uniquely integrating quantitative predictive modeling with qualitative generative AI diagnostics. The system is built on two core components: first, a RUL Prediction Model that uses the XGBoost algorithm, enhanced with Statistical Process Control (SPC) to filter 148,020 sensor records from a PostgreSQL database, predicting the Remaining Useful Life (RUL). Second, a Technician Assistant that employs the Mistral 7B Large Language Model (LLM), integrated with Retrieval-Augmented Generation (RAG), to provide natural language answers to technician queries by accessing relevant maintenance logs. The RUL model demonstrated high performance with an R^2 of 0.90, RMSE of 246.41 hours, and MAE of 173.59 hours, identifying machine model and age as key predictors. While 8,334 outliers highlighted data noise challenges, the SPC enhancement proved crucial for reliability. The LLM assistant successfully delivered targeted, relevant maintenance insights, enhancing decision-making. This combined system significantly optimizes maintenance schedules and reduces operational disruptions, demonstrating a strong potential for real-time industrial deployment.

Keywords**:** Preventive Maintenance, RUL, LLM, XGBOOST, RAG, SPC, MISTRAL 7B

ABSTRAK

Projek berasaskan penyelidikan ini bertujuan untuk mengurangkan masa henti tidak dirancang dan meningkatkan kecekapan penyelenggaraan dalam persekitaran industri, didorong oleh penalti kewangan dan operasi yang ketara yang berkaitan dengan kegagalan peralatan yang tidak dijangka. Kebaruannya adalah menyepadukan pemodelan ramalan kuantitatif dengan diagnostik kecerdasan buatan generatif kualitatif bagi menghasilkan pandangan boleh dilaksana. Sistem ini dibina di atas dua komponen teras: pertama, Model Ramalan RUL yang menggunakan algoritma XGBoost, dipertingkatkan dengan Kawalan Proses Statistik (SPC) untuk menapis 148,020 rekod sensor daripada pangkalan data PostgreSQL, meramalkan Jangka Hayat Berguna Baki (RUL). Kedua, Pembantu Juruteknik yang menggunakan Model Bahasa Besar (LLM) Mistral 7B, disepadukan dengan Penjanaan Dipertingkat-Perolehan (RAG), untuk memberikan jawapan bahasa semula jadi kepada pertanyaan juruteknik dengan mengakses log penyelenggaraan yang relevan. Model RUL menunjukkan prestasi tinggi dengan R^2 0.90, RMSE 246.41 jam, dan MAE 173.59 jam, mengenal pasti model mesin dan umur sebagai peramal utama. Walaupun 8,334 nilai terpencil menyerlahkan cabaran hingar data, penambahbaikan SPC terbukti penting untuk kebolehpercayaan. Pembantu LLM berjaya menyampaikan pandangan penyelenggaraan yang disasarkan dan relevan, meningkatkan pembuatan keputusan. Gabungan sistem ini secara signifikan mengoptimumkan jadual penyelenggaraan dan mengurangkan gangguan operasi, menunjukkan potensi kukuh untuk penggunaan industri masa nyata.

Kata Kunci: Penyelenggaraan Pencegahan, RUL, LLM, XGBOOST, RAG, SPC, MISTRAL 7B

INTRODUCTION

In the domain of industrial operations, the continuity of manufacturing processes is critical, yet frequently undermined by sudden equipment failures. These disruptions lead to significant production declines and incur substantial financial and operational penalties for industries. To mitigate these risks, industries have historically relied on preventive maintenance (PM), a strategy centred on estimating the Remaining Useful Life (RUL) of machinery to anticipate breakdowns before they occur (Motrani et al., 2024). Traditionally, RUL estimation and maintenance scheduling have been manual processes, heavily dependent on the statistical intuition and domain expertise of technicians. However, this human-centric approach is often insufficient; modern industrial machinery involves complex, non-linear variables that make manual prediction labour-intensive and prone to error (Al-Refaie et al., 2025). Furthermore, when failures do occur, technicians must often navigate fragmented manual logs and guides to diagnose root causes, a process that prolongs downtime and reduces maintenance efficiency.

The emergence of Artificial Intelligence (AI) and Generative AI offers a transformative opportunity to bridge the gap between complex sensor data and actionable maintenance decisions (IAS Research, 2025). While traditional methods struggle with high-dimensional data, AI-driven systems can process vast datasets to uncover hidden patterns, offering a path to "predictive" rather than merely "preventive" maintenance. This project leverages this technological shift to develop a dual-component system designed to empower technicians with data-driven insights. By integrating advanced machine learning prognostics with Large Language Model (LLM)-aided design, the system addresses the two primary bottlenecks of industrial maintenance: the inaccuracy of failure prediction and the latency in decision-making. The proposed solution not only predicts when a machine will fail but also assists technicians in understanding why, utilizing Retrieval-Augmented Generation (RAG) to contextualize data within existing maintenance knowledge bases.

To realize this system, this research implements a robust technical architecture comprising a predictive model and an intelligent assistant. The RUL prediction component utilizes the XGBoost gradient-boosting algorithm, enhanced with Statistical Process Control (SPC) to effectively filter noise and outliers from a dataset of 148,020 sensor records stored in PostgreSQL. This enhancement is critical for handling the stochastic nature of industrial data, ensuring reliable prognostics (Soualhi et al., 2024); notably, the model achieves an R^2 of 0.90, demonstrating high predictive accuracy. Complementing this, the technician assistant is powered by the Mistral 7B LLM integrated with RAG architecture. This component allows the system to query unstructured maintenance logs and provide natural language responses to complex queries, such as identifying failure causes. By synergizing high-precision RUL estimation with an AI agent capable of identifying log anomalies, this system significantly optimizes maintenance schedules and reduces operational disruptions. Unlike existing isolated maintenance solutions, the novelty of this work resides in its explicit coupling of statistical prognostics with generative AI diagnostics. By ensuring the LLM is tightly constrained by verified maintenance logs via the RAG architecture, the system overcomes the limitations of purely quantitative prediction models and mitigates the hallucination risks typical of standalone generative AI in industrial settings.

EXECUTIVE SUMMARY

This project follows the CRISP-DM methodology to develop a dual-component system: an XGBoost RUL prediction model enhanced with SPC and a Mistral 7B RAG assistant. The process began with Data Acquisition, retrieving 800,000 telemetry records (voltage, rotation, temperature, pressure) from 100 machines via an Azure Dataset, which were stored in a local PostgreSQL database and accessed using SQLAlchemy. Data Preprocessing involved converting timestamps, merging tables to define the RUL target (next failure), filtering RUL (0-4000h), and applying a square root transformation and StandardScaler normalization for the RUL model. For the LLM assistant, maintenance and error logs were extracted as kept in the local database, and embeddings were generated using SentenceTransformer to support RAG retrieval. Feature Engineering created over 45 features for RUL prediction, including temporal features (e.g., time since last maintenance), rolling means (6, 12, and 24-hour windows), and binary flags for SPC outliers (values beyond 3 standard deviations). Model Training involved an 80/20 stratified train-test split, using grid search to optimize the XGBoost model with regularization, early stopping, and sample weights. The Mistral 7B model was fine-tuned using domain-specific prompts and RAG (via cosine similarity) with a temperature of 0.5. The fully-developed dual-component system was deployed into a usable web-based dashboard, supported by Telegram API for

mobile notifications. The fishbone diagram as shown in Figure 1 illustrates the holistic integration of the system through its operational dimensions which are Measurement, Algorithms, Technology and Man which converge to achieve the primary objectives of mitigated downtime and improved efficiency. Within the Measurement domain, a robust data foundation is established by combining raw machine telemetry and a consolidated knowledge base with Statistical Process Control (SPC) for quality assurance, while the Algorithms dimension drives the computational logic using an XGBoost regressor for RUL prediction validated by rigorous performance metrics.

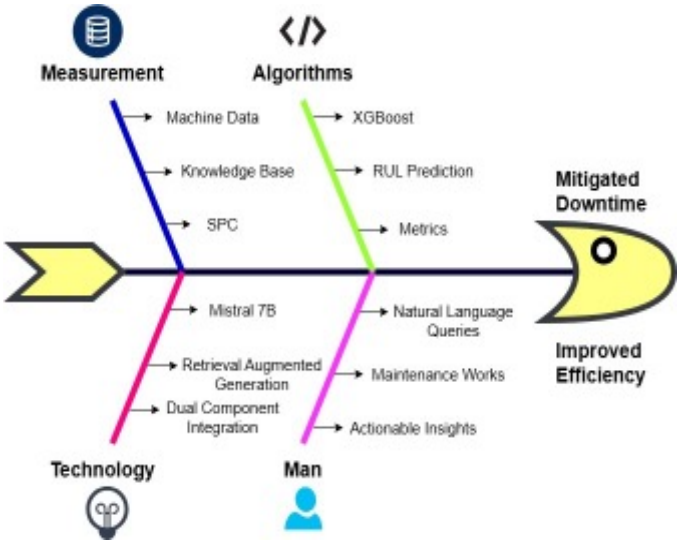

FIGURE 1. Fishbone Diagram of the project

Simultaneously, the Technology sector defines the generative AI architecture through the deployment of Mistral 7B within a Retrieval-Augmented Generation (RAG) framework to ensure the seamless Dual Component Integration of numerical predictions with the intelligent assistant. Finally, the Man dimension encapsulates the human-in-the-loop workflow by translating technicians' natural language queries into actionable insights for precise maintenance works, ultimately resulting in the reduction of unplanned stoppages and the optimization of industrial operations.

OBJECTIVES

This project aimed to build and evaluate a dual-component preventive maintenance system for generic industrial machines. The primary objectives were to create a robust XGBoost model to predict RUL (Remaining Useful Life) and to develop a Mistral 7B LLM assistant with RAG for technician support. The RUL model's reliability was enhanced using Statistical Process Control (SPC) to manage sensor data outliers. The system was developed using a PostgreSQL (AzureTPMDB) dataset containing telemetry (like temperature, voltage, rotation, and pressure), failure logs, and maintenance records. The implementation relied on Python libraries such as SQLAlchemy, pandas, and scikit-learn. The scope covered the full data science workflow, including preprocessing, feature engineering, training, and evaluation. Key limitations of the project include its focus on generic rather than specific machines and the qualitative evaluation of the LLM assistant due to a lack of specific metrics. The project also excluded hardware-specific optimizations and non-industrial applications.

METHODOLOGY

This project adopts the *Cross-Industry Standard Process for Data Mining (CRISP-DM)* framework. This structured approach ensures a rigorous lifecycle for developing the dual-component system: the XGBoost-based Remaining Useful Life (RUL) predictor and the Mistral 7B-based intelligent technician assistant as depicted in Figure 2.

FIGURE 2. CRISP-DM framework flow from CRISP-DM Process Diagram, by K. Jensen, 2012, Wikimedia Commons

Business Understanding

The initial phase of the methodology focuses on addressing the critical operational challenges within industrial settings, specifically the significant financial and operational penalties associated with unexpected equipment failure. The primary business objective is to mitigate unplanned downtime and enhance maintenance efficiency by transitioning from reactive protocols to a predictive maintenance strategy. To achieve this, the project identifies the need for a system that empowers technicians with actionable, data-driven insights, thereby optimizing maintenance schedules and reducing operational disruptions. These business objectives translate into two distinct technical requirements: (1) the development of a high-precision Remaining Useful Life (RUL) prediction capability to anticipate failures before they occur, and (2) the integration of an intelligent knowledge retrieval system to facilitate rapid decision-making. By defining these goals, the project establishes the necessity for a dual-component architecture combining XGBoost-based regression with a Mistral 7B-powered Large Language Model (LLM) to bridge the gap between raw sensor data and practical maintenance execution.

Data Understanding

The project utilizes the Azure Telemetry Dataset, comprising approximately 800,000 telemetry records from 100 generic industrial machines. Table 1 shows the dataset, which includes high-frequency sensor readings (voltage, rotation, pressure, vibration), failure logs, maintenance history, and machine metadata. To simulate a robust industrial data pipeline, the raw data was ingested into a PostgreSQL database. Data retrieval and manipulation were managed using Python's SQLAlchemy library, ensuring efficient querying of the temporal and relational data required for both predictive modeling and context retrieval.

TABLE 1. Sample from utilized Microsoft Azure Dataset

datetime	machineID	volt	rotate	pressure	vibration
1/01/2015 6:00	1	176.217	418.504	113.077	45.087
1/01/2015 7:00	1	162.879	402.747	95.460	43.413
1/09/2015 21:00	2	191.749	522.020	101.408	43.2966
1/09/2015 22:00	2	158.541	439.610	94.801	46.547

Data Preparation

Preparation focused on preprocessing, ensuring data consistency and establishing distinct pipelines for the numerical and textual components. To support RUL prediction, timestamps across telemetry, failure, and maintenance tables were temporally aligned to define the target variable calculated as the time to the next failure capped at 4,000 hours after which a square root transformation was applied to stabilize target variance and sensor features were normalized using StandardScaler. In parallel, to enable the Intelligent Assistant, textual data comprising maintenance logs and error reports was consolidated into a knowledge base and processed using a SentenceTransformer model to generate high-dimensional vector embeddings, facilitating efficient semantic retrieval. To capture complex degradation patterns, extensive feature engineering was performed to generate over 45 input variables, integrating temporal metrics like time since last maintenance with rolling means and standard deviations calculated over 6, 12, and 24-hour windows to track short- and medium-term operational trends. Additionally, specific Statistical Process Control (SPC) measures were implemented to distinguish noise from genuine anomalies by flagging sensor values deviating beyond three standard deviations via binary indicators, while machine metadata was one-hot encoded to account for hardware-specific variance.

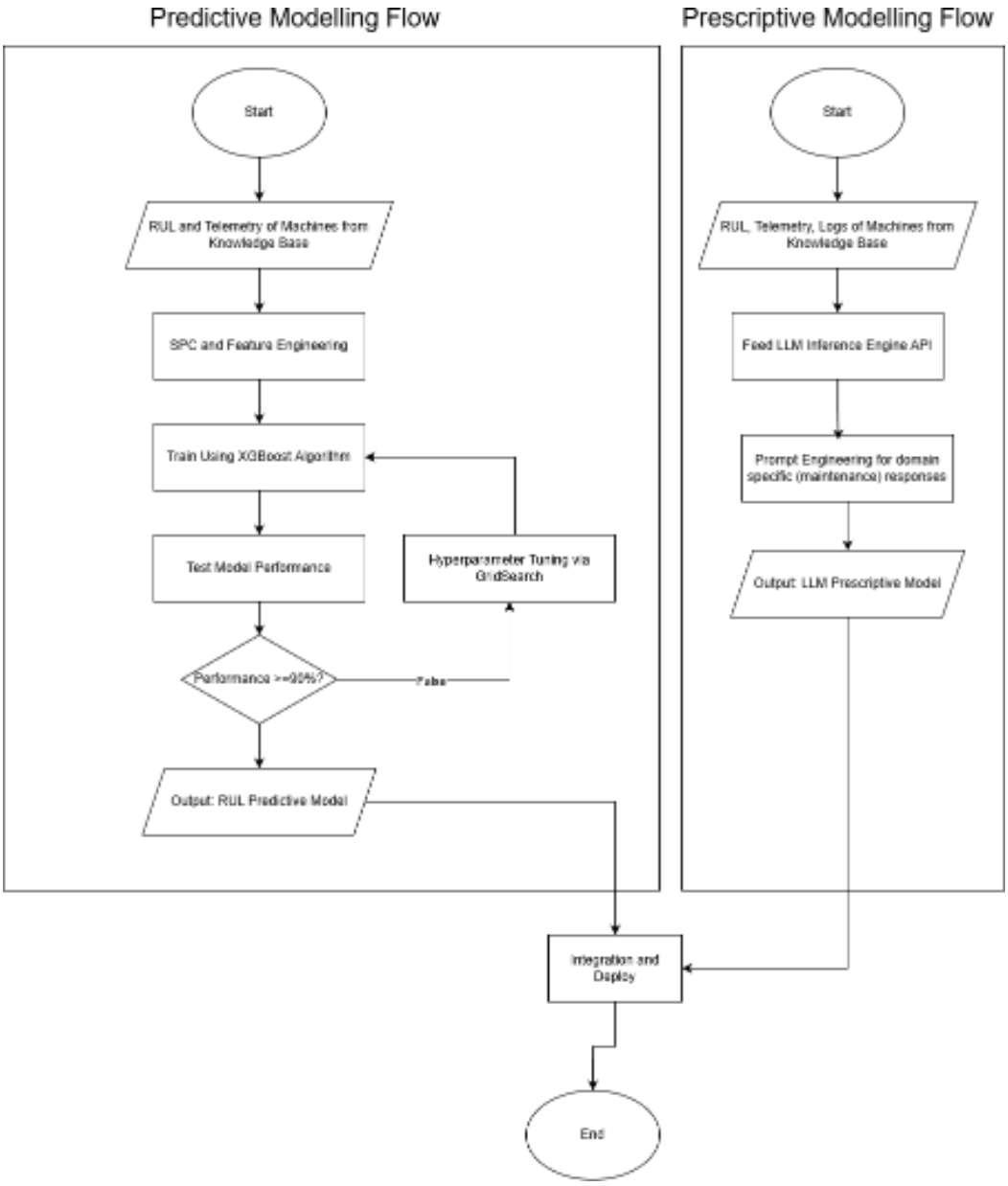

FIGURE 3. Modelling flowchart for Predictive and Prescriptive Models

Modellng

The implementation phase executed the development of the dual-component system, beginning with an XGBoost Regressor for RUL prediction trained on an 80/20 dataset split stratified by RUL quartiles to ensure balanced failure stage distribution. Hyperparameters were optimized via GridSearchCV, while sample weights inversely proportional to RUL magnitude were applied to prioritize learning from critical, near-failure states. Complementing this, the technician assistant utilizes a Retrieval-Augmented Generation (RAG) framework where vectorized queries retrieve relevant logs via cosine similarity; these logs provide context for a Mistral 7B Large Language Model (temperature 0.5), enabling the generation of factually grounded, domain-specific responses without the need for full fine-tuning.

Evaluation

To validate system performance, a rigorous dual-evaluation strategy was employed. The predictive capabilities were assessed quantitatively using several regression metrics; Simultaneously, the Intelligent Assistant underwent qualitative validation via "LLM-as-a-Judge" methodology to score relevance and coherence, ensuring the generation of safe, actionable technical advice. To rigorously assess the predictive accuracy of the model, the project employs the Root Mean Squared Error (RMSE) metric, which is calculated using the formula (1).

(1)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE was selected for its ability to heavily penalize larger deviations that could result in critical unplanned downtime. In the context of this specific system, y_i represents the ground truth Remaining Useful Life (RUL) derived directly from the dataset, while $\hat{y}_i$ denotes the corresponding RUL value estimated by the XGBoost regressor for each specific sensor reading. Additionally, n refers to the total number of sensor observations within the test dataset, and the term $(y_i - \hat{y}_i)$ quantifies the temporal error between the predicted failure time and the actual operational limit of the machinery. To provide a clear representation of the average error magnitude without the sensitivity to outliers inherent in RMSE, the project utilizes the Mean Absolute Error (MAE), calculated as using the formula (2).

(2)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Within this predictive framework, y_i corresponds to the actual Remaining Useful Life extracted from maintenance logs, while $\hat{y}_i$ represents the RUL predicted by the XGBoost model for the i^{th} sensor observation among the n records in the validation set. Crucially, the term $y_i - \hat{y}_i$computes the absolute difference between these values, ensuring that the direction of the error whether the model overestimates or underestimates the failure time does not skew the assessment of the model's general reliability. To evaluate the goodness of fit and determine the proportion of variance in the Remaining Useful Life that is predictable from the sensor telemetry, the study calculates the Coefficient of Determination R^2 using the formula (3)

(3)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In this expression, the numerator represents the sum of squared residuals capturing the deviation between the actual RUL y_i and the XGBoost-predicted RUL y_i while the denominator denotes the total sum of squares, measuring the variance of the observed data relative to the mean RUL $\bar{y}$. This metric serves as a critical indicator of the model's explanatory power, quantifying its improvement over a baseline mean-based prediction model.

For the LLM assistant, evaluation is qualitative assessment, focusing on the relevance, clarity, which are derived from the retrieved context and natural language processing capabilities of Mistral 7B, ensuring the solution meets practical requirements before deployment as shown in Table 2.

TABLE 2. Criteria to evaluate the prescriptive Large Language Model

Judging Criteria	Description
Relevance	Measuring the extent to which the response directly addressed the user's query.
Actionability	Assessing how effectively the response provided information that could guide technicians in determining the next steps.
Factual Accuracy	Verifying whether the response aligned with established maintenance guidelines such as (ISO 17359, NIOSH) and domain-specific knowledge.
Coherence	Evaluating the logical consistency of the response within the context of the technical domain.

RESULTS AND DISCUSSION

Predictive Model

The performance of the predictive model was assessed using regression metrics (1, 2, 3) as stated in the evaluation part of the methodology. On the original scale as shown in Table 3, the model achieved an RMSE of 242.68 hours and MAE of 170.72 hours, with an R^2 of 0.90, indicating that 90% of the variance in RUL was successfully explained. From an industrial perspective, these results suggest that the model is highly reliable for predictive maintenance planning. For instance, an average absolute error of 170 hours (roughly one week of operation in many industrial settings) provides sufficient lead time for maintenance teams to schedule interventions, order spare parts, and minimize unplanned downtime.

TABLE 3. Regression based evaluation for RUL Predictive Model

Model Name	MAE	RMSE	R^2	RUL Horizon
XGBoostV2_Postgre	170.7h	242.7h	0.9	Up to 4000h

Figure 3 shows the graph of Predicted vs. Actual RUL, revealing a heavy data concentration in the lower-left quadrant that reflects a realistic maintenance imbalance where "healthy" states outnumber failures. The model demonstrates peak reliability in the critical zone (0–1000h), where the highest sample density (*N*=102,405) aligns with an empirical coverage probability of 98.4%, indicating nearly all predictions fall within the 95% confidence bounds. However, as the prediction horizon extends, predictive certainty degrades: coverage drops to 88.4% in the 1000–2000h segment (*N*=33,348) and further to 71.3% in the 2000–3000h range. In the longest-term forecast (3000–4000h), coverage reaches its minimum at 68.0% (*N*=3,533), with scatter points exhibiting a distinct tendency to drift below the ideal trajectory.

The results highlight a model that is inherently risk-averse and highly calibrated for immediate operational decision-making, validating the sample weighting strategy which successfully minimized variance in the critical 0–1000h window. This approach aligns with the machinery health prognostic framework established by Lei et al. (2018), who emphasize that industrial cost functions must prioritize imminent failure detection over long-term linearity. The observed degradation in coverage for long-term predictions corresponds to the theoretical challenges of aleatoric uncertainty, where the accumulation of stochastic sensor noise naturally erodes confidence over extended horizons (Soualhi et al., 2024). Notably, the conservative drift below the ideal line in the 3000–4000h segment provides a preferable industrial safety margin, avoiding false security without triggering premature maintenance. This specific distribution explains the divergence in the quantitative metrics: the high R^2 (0.90) and low MAE (173.59h) are driven by the dense, accurate clustering in the immediate action zone, while the larger deviations in the sparse, long-term data disproportionately inflate the RMSE to 246.41 hours (Al-Refaie et al., 2025).

The graph shown in Figure 4 plots Prediction Error against Actual Remaining Useful Life (RUL), revealing a distinct heteroscedastic pattern where residual variance expands as the target variable increases. In the critical near-failure phase (0–1000h), residuals converge tightly around the zero-error line, exhibiting the highest model stability despite a maximum recorded deviation of 1898 hours. As the

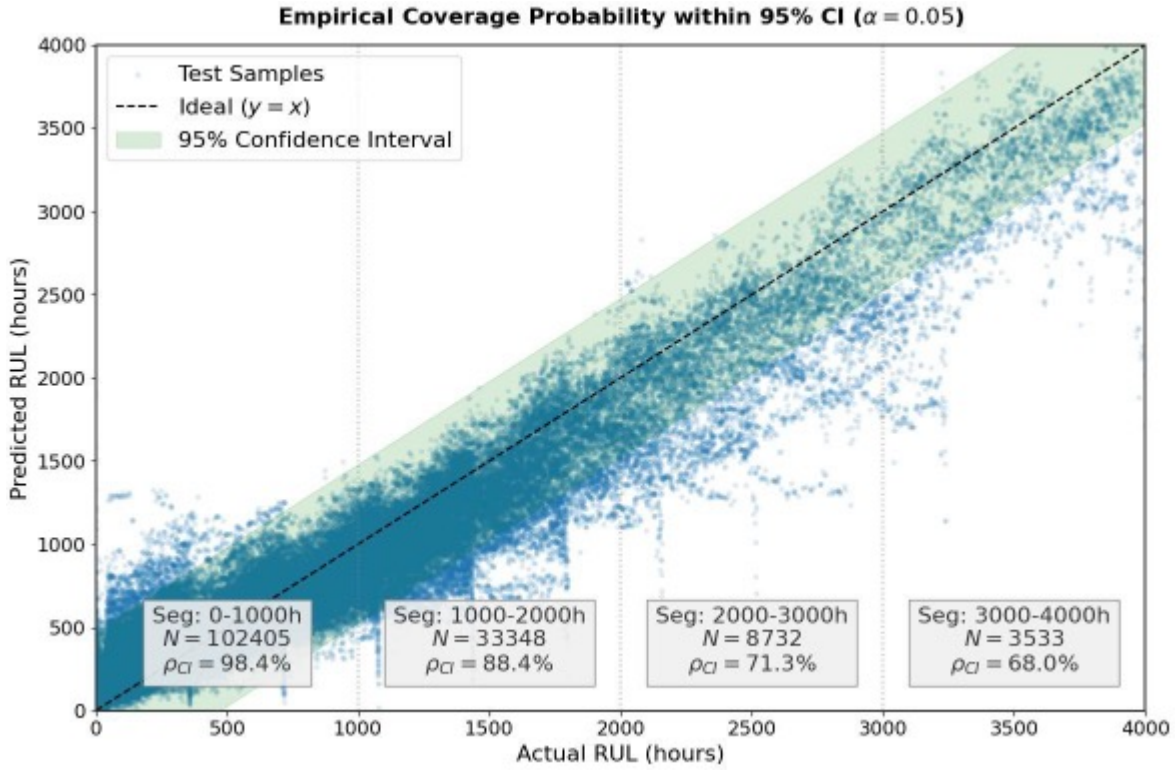

FIGURE 4. Predictive model predictions within 95% confidence interval

prediction horizon extends, dispersion progressively widens: maximum deviations fluctuate from 1597 hours in the 1000 – 2000h segment to 2057 hours in the 2000 – 3000h range, eventually peaking at 2292 hours in the early-life 3000 – 4000h segment. This trend indicates that the model is most confident and consistent when the machine is approaching failure, with uncertainty naturally accumulating as the operational state moves further from the failure event.

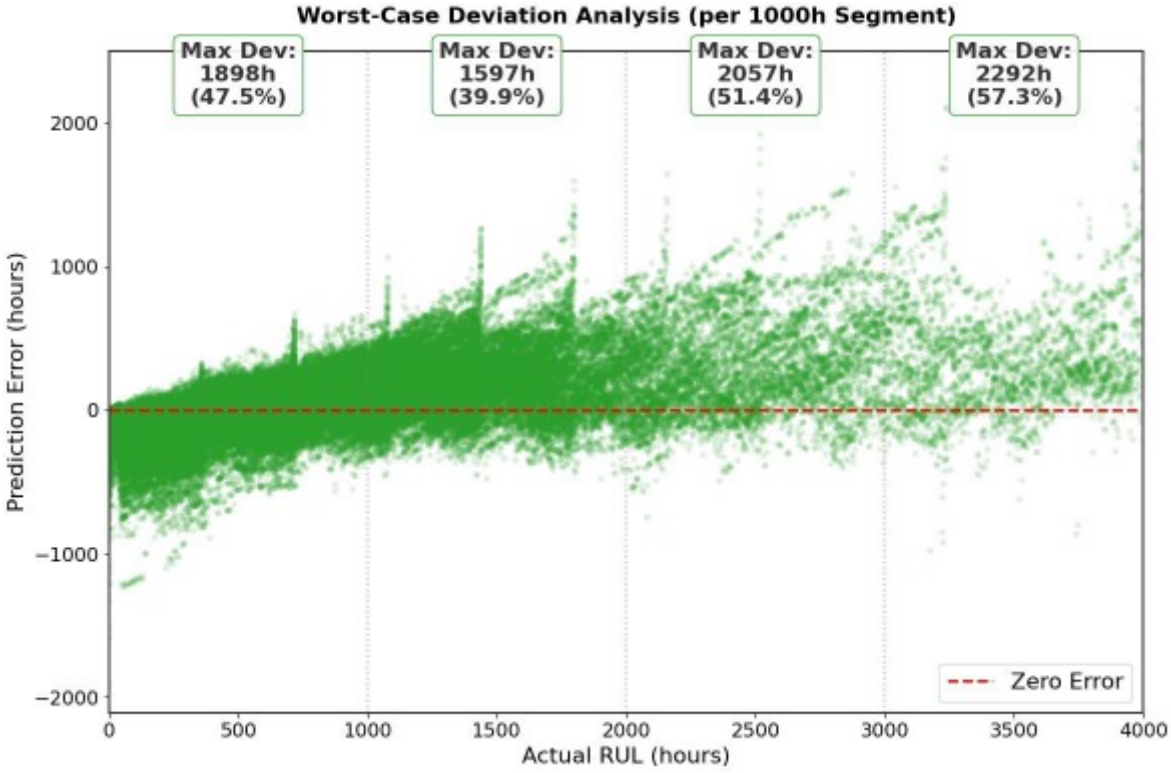

FIGURE 5. Predictive model prediction residuals

The residual analysis corroborates that the model successfully aligns with maximizing prognostic accuracy specifically within the actionable maintenance window. This behavior is a direct, engineered outcome of the training strategy; by assigning weights inversely proportional to RUL magnitude, the loss function penalizes errors in the short-term (critical) range significantly more heavily than errors in the long-term (healthy) range. This methodology effectively addresses the non-linear feature reduction challenges, ensuring that limited computational capacity is focused on high-stakes predictions. The observed heteroscedasticity highlights the "prediction horizon dilemma" common in regression-based prognostics. As noted by Wang et al. (2023), as the prediction horizon extends (here, beyond 3000 hours), variable future operating conditions make precise determination of the "Time to Failure" statistically indistinguishable from noise. However, from an industrial standpoint, this variance is acceptable; knowing a machine is in a "healthy" state is sufficient for planning, whereas precise quantification is only required as the asset approaches the failure threshold (Motrani et al., 2024).

The distribution shown in Figure 5 contrasts the frequency density of residuals against a theoretical Gaussian baseline, revealing a symmetrical distribution centred almost perfectly at zero that indicates negligible systematic bias. Structurally, the distribution exhibits a distinct leptokurtic profile: the histogram peak reaches a density of approximately 0.0025 significantly exceeding the theoretical normal peak of 0.0016 demonstrating that the vast majority of errors are tightly clustered within the - 500 to +500 hour range. While the core data mass is highly concentrated, thin tails extending outward to ±2000 hours persist, confirming the presence of the rare but high-magnitude outliers previously identified in the scatter plot analysis.

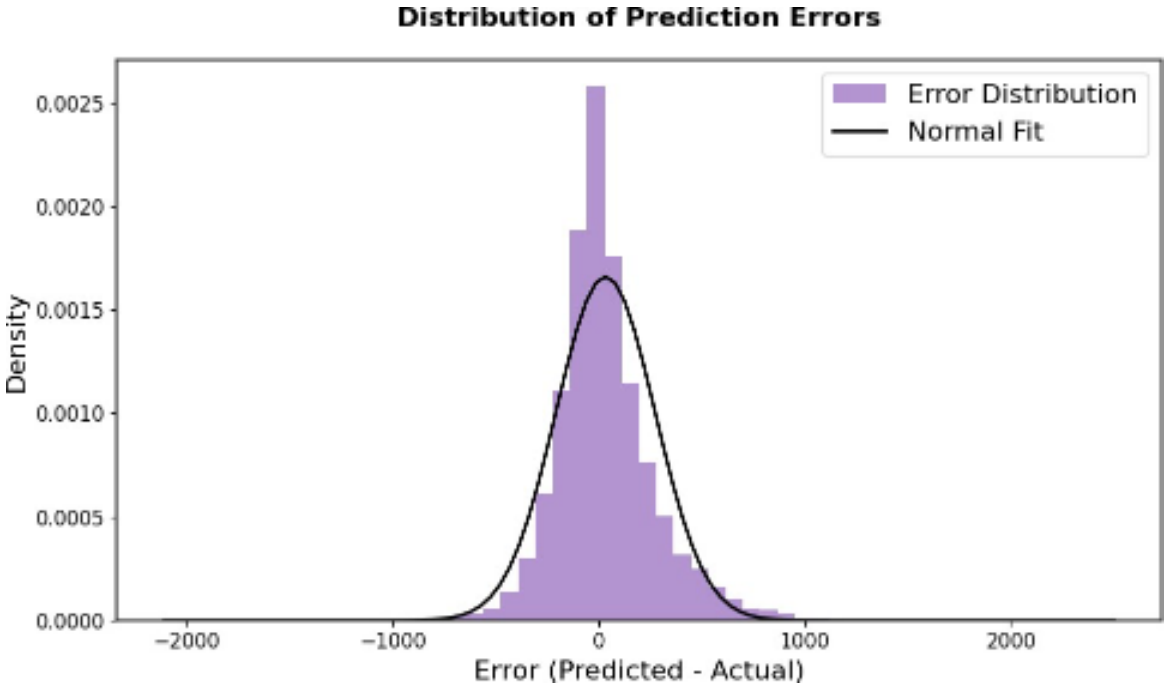

FIGURE 6. Distribution of errors in prediction

The histogram serves as a statistical validation of the model's neutrality, as the symmetry around the zero-line demonstrates that the algorithm balances its predictions without systematic over-or under-estimation of machine life. The observed high kurtosis is particularly significant; statistically, it indicates that the error distribution is leptokurtic. This implies that for the majority of instances, the model has successfully learned the deterministic degradation signals, leaving only residuals that are significantly smaller than those expected from random Gaussian noise. This distributional shape offers the mathematical bridge between the conflicting error metrics: the dense central clustering drives the low MAE and high R^2, rewarding consistent accuracy. Conversely, the extended, low-frequency tails correspond to the divergence of the RMSE. Because RMSE involves squaring residuals, these distal outliers exert a disproportionately heavy penalty, effectively serving as a risk-sensitive gauge for worst-case scenarios in safety-critical applications (Rezende et al., 2022).

Prescriptive LLM

Figure 6 presents the bar chart illustrating the aggregate effectiveness of the Mistral 7B technician assistant across four distinct evaluation metrics, with bar heights representing mean scores on a 1–10 scale and error bars denoting 95% Confidence Intervals. Relevance and Coherence achieved joint-highest mean scores of 8.74, exhibiting narrow confidence intervals that signal high consistency in both log retrieval and linguistic structure. Factual Accuracy followed closely at 8.56 with overlapping error margins, whereas Actionability recorded the lowest mean at 7.70, characterized by a slightly wider dispersion compared to the other metrics.

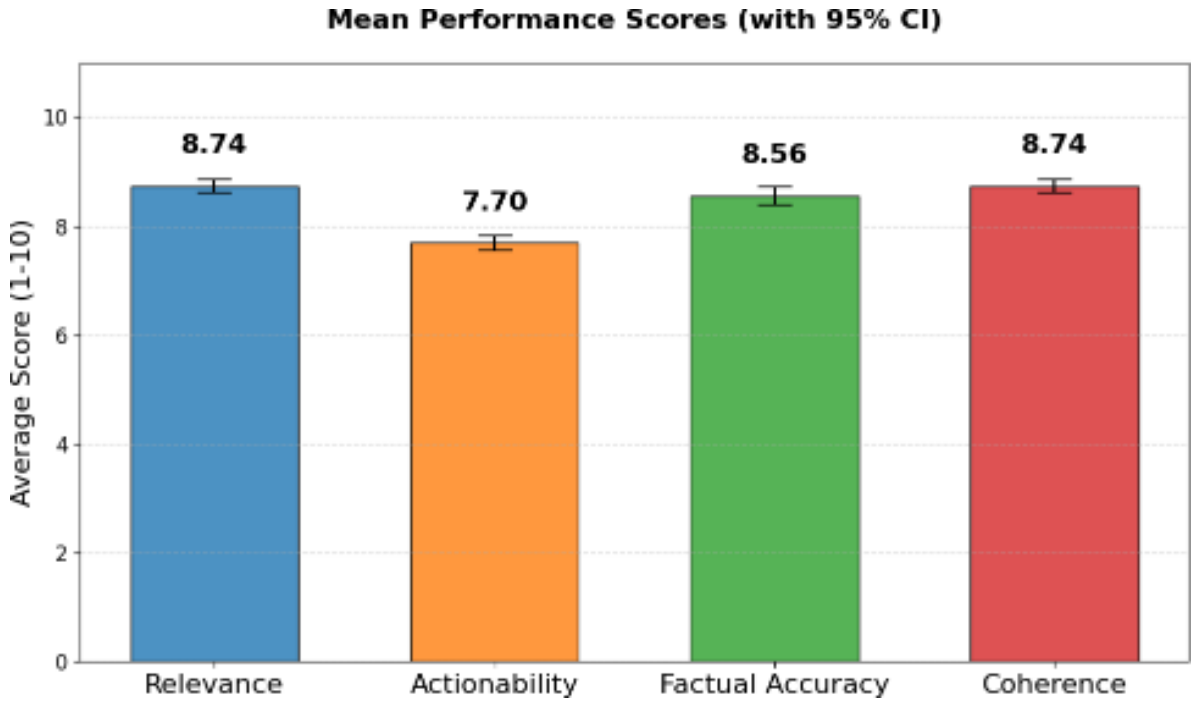

FIGURE 7. Performance score of LLM prescription and diagnostics

The results empirically validate the RAG architecture, where high Relevance (8.74) and Factual Accuracy (8.56) confirm that the embedding mechanism effectively retrieved context that the LLM utilized faithfully. This supports the foundational premise by Lewis et al. (2020) that retrieving external non-parametric knowledge significantly mitigates hallucination risks in generative models. However, the distinct deviation in Actionability (7.70) reflects the inherent constraints of deploying LLMs in industrial operations. While the model excels at diagnostic explanation, its conservative approach to recommending physical repairs is a predictable byproduct of safety-aligned training. As observed in similar implementations by Harbola et al. (2025) and Löwenmark et al. (2025), agent-based systems in maintenance often suppress potentially hazardous interventions in the absence of explicit procedural source text. This limitation is further supported by Zhang et al. (2024), who note that while RAG frameworks improve information retrieval in IT and industrial operations, they require rigid guardrails to ensure that prescriptive advice does not exceed the verified safety protocols contained within the vector database.

CONCLUSION

The comparative analysis of the predictive (RUL) and prescriptive (LLM Assistant) components reveals a unified system characteristic: a distinct operational conservatism suited for safety-critical environments. For the predictive component, results demonstrate a model highly calibrated for immediate decision-making, where the residual analysis and empirical coverage probability highlight a deliberate trade-off engineered through sample weighting. This strategy maximized accuracy in the critical 0–1000h window (R^2 0.90, MAE 173.59h) at the expense of long-term precision, a behavior that aligns with the theoretical constraints of aleatoric uncertainty and the statistical impossibility of precise long-term prediction in

stochastic environments. Crucially, the high kurtosis observed in the error histogram validates the model's reliability, proving it can distinguish deterministic degradation signals from random noise with high confidence. This risk-averse behavior is mirrored in the prescriptive LLM component; while high Relevance (8.74) and Factual Accuracy (8.56) scores confirm the efficacy of the Retrieval-Augmented Generation (RAG) architecture, the lower Actionability (7.70) score reflects a safety-aligned limitation. Just as the RUL model is conservative in predicting long-term health, the Mistral 7B assistant effectively suppresses potentially hazardous interventions when explicit procedural text is absent, ensuring the system functions as a safe decision-support tool rather than an autonomous operator.

This project successfully achieved its primary objective of developing a dual-component preventive maintenance system for generic industrial machines, encompassing the full data science workflow from preprocessing to evaluation. The goal of creating a robust RUL predictor was realized through an XGBoost algorithm enhanced by Statistical Process Control (SPC), which effectively filtered outliers from the AzureTPMDB dataset comprising telemetry, failure logs, and maintenance records to deliver a model that prioritizes actionable maintenance windows. Simultaneously, the objective to support technicians was met by developing a Mistral 7B assistant integrated with RAG, leveraging SQLAlchemy and SentenceTransformers to transform raw PostgreSQL logs into natural language insights that bridge the gap between complex data and technician understanding. While the study acknowledges limitations specifically the exclusion of hardware-specific optimizations like vibration signature analysis and the reliance on qualitative evaluation for the LLM due to the lack of standardized generative AI metrics the system demonstrates a robust, end-to-end framework for minimizing unplanned downtime and enhancing maintenance efficiency in industrial settings. Ultimately, the primary novelty of this study is the successful operationalization of a hybrid predictive-prescriptive framework. It offers an end-to-end solution that bridges rigorous machine learning with nuanced natural language processing, establishing a new standard for human-in-the-loop industrial maintenance systems.

ACKNOWLEDGEMENTS

The authors would like to express deepest gratitude to Dr. Siti Azirah Binti Asmai and Dr. Zuraida binti Abal Abas for their expert guidance. A sincere appreciation is also extended to Ministry of Higher Education (KPT), Universiti Teknikal Malaysia Melaka (UTeM), Centre for Student Empowerment, Alumni & Careers (SPAC), and Centre for Enterprise and Technopreneurship Development (CREATE), Office of Deputy Vice Chancellor (Student Affairs & Alumni) for the valuable opportunity, support and sponsorship.

REFERENCES

Al-Refaie, A., et al. (2025). Prediction of the remaining useful life of a milling machine using machine learning. *MethodsX*, *12*, 102500.

Bourdin, M., et al. (2025). An agile method for implementing RAG tools in industrial SMEs. *arXiv preprint arXiv:2502.12345*.

Harbola, A., et al. (2025). Prescriptive agents based on RAG for automated maintenance (PARAM). *arXiv preprint arXiv:2501.01234*.

IAS Research. (2025). Generative AI and retrieval-augmented generation (RAG) in electrical and computer engineering. *IAS Research Journal*, *4*(2), 12–18.

Jiang, A. Q., et al. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Khan, S., & Yairi, T. (2018). A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, *107*, 241–265.

Lei, Y., Li, N., Guo, L., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Löwenmark, K., et al. (2025). Agent-based condition monitoring assistance via multimodal RAG. *arXiv preprint arXiv:2502.05678*.

Motrani, M., et al. (2024). Predictive maintenance: A machine learning approach to remaining useful life (RUL) estimation of industrial machines. *Journal of Intelligent Manufacturing*, *35*(2), 45–60.

Pan, A., et al. (2023). RAGLog: Log anomaly detection using retrieval augmented generation. *arXiv preprint arXiv:2311.02564*.

Rezende, J. G., de Oliviera, M., & Carvalho, J. (2022). A review of RUL estimation metrics in the context of industrial prognosis. *International Journal of*

Prognostics and Health Management, *13*(1).

Soualhi, A., et al. (2024). Explainable RUL estimation of turbofan engines based on prognostic indicators and heterogeneous ensemble machine learning predictors. *Engineering Applications of Artificial Intelligence*, *128*, 107380.

Wang, A., et al. (2023). Remaining useful life prediction of aircraft turbofan engine based on random forest feature selection and multi-layer perceptron. *Applied Sciences*, *13*(14), 8205.

Zhang, T., et al. (2024). RAG4ITOps: A supervised fine-tunable RAG framework for IT operations and maintenance. *arXiv preprint arXiv:2401.00256*.

Suthareesanan A/L Saravanan*
No.37, Lorong Bukit Minyak 25,
Taman Bukit Minyak Indah,
14000 Bukit Mertajam, Penang, Malaysia.

Siti Azirah Asmai & Zuraida Abal Abas
Faculty of Artificial Intelligence and Cyber Security,
Universiti Teknologi Melaka (UTeM),
76100 Durian Tunggal, Melaka, Malaysia.

*Corresponding author: B032320096@student.utem.edu.my