ProphetJet: Predictive Maintenance Modelling Using LSTM, Random Forest, and XGBoosting to Forecast RUL Metrics of NASA Turbofan Jet Engines

Arjan Waraich
University of Toronto Schools
waraicharjan97@gmail.com

Max Huddleston
University of Toronto Schools
codr.9595@gmail.com

Kushad Manikandan
University of Toronto Schools
kmkushad@gmail.com

Sidney Shu
University of Toronto Schools
shusi@utschools.

Dora Li
University of Toronto Schools
lido@utschools.ca

Jaotin Ling
University of Toronto Schools
linja@utschools.ca

Abstract*—This project develops a predictive maintenance model for jet engines using the NASA C-MAPSS dataset. The model utilizes supervised learning to classify engine health states and predict Remaining Useful Life (RUL). Key techniques include data preprocessing, feature engineering, and machine learning algorithms optimized for time-series forecasting. Model performance is evaluated using RMSE, MAE, and overall loss between epoch gradients, with correlation matrices aiding feature selection. Future improvements include advanced deep learning techniques to enhance accuracy and adaptability, allowing machine owners to fine-tune the model with custom data for broader deployment. The model achieves a considerable accuracy of 87.4% with a *2%* standard deviation. This approach enables proactive maintenance, reducing downtime and operational costs. *See the project Github.

I. Introduction

In modern machine and industrial operations, especially in industries such as manufacturing, warehousing, and aerospace, the reliability of complex machinery and robotics is critical in ensuring both safety, efficiency, cost-effectiveness, and coherence. The practice of *Predictive Maintenance* has emerged as a novel strategy for mitigating unplanned failures of machines, detecting anomalies, result optimization, safety regulation, and business cost cutting – by leveraging data-driven techniques and sensor technologies, such as **IoT** or **LPWAN** (Low-Power Wide-Area Networks) etc., to anticipate and predict equipment degradation before critical faults occur. Unlike reactive maintenance, which addresses failures after they happen, predictive maintenance enables logistically tactical proactive interventions, which help fine-tune maintenance schedules, reducing operational disruptions – directly improving cost efficiency in terms of pay for routinely-established repair tasks, for example. More specifically, in aerospace and related industrial settings, the ability to accurately predict the Remaining Useful Life (RUL) of jet engines is critical in saving time, lives, and costs.

A. Motivation

In industry, RUL is a key metric defined as the estimated time an asset or component has left before it needs to be replaced or repaired, making it key information for predictive maintenance and asset optimization. By forecasting the number of cycles an engine can operate before requiring maintenance or failure – at a certain operational setting or under critical circumstances of key variables (for example @ 78% engine power at an altitude of 32,000 feet, with a certain EDR turbulence setting, under certain external wind conditions) – the benefits of the implications of accurate RUL prediction extends to minimizing unexpected failures, enhancing safety (commercial), and saving both *time* and *money*. [1] However, conventional approaches to RUL estimation struggle heavily with real-world conditions due to the complexity of engine degradation and the variability in operational environments, the nature of which is by virtue of the tremendous amount of data generated by sensors on a time-series basis. Due to the inept nature of traditional methods, certain deep learning techniques, particularly those capable of analyzing time-series data from engine sensors, enable more precise forecasting of potential failures. For instance, Long Short-Term Memory (LSTM) networks have been effectively utilized to predict RUL by learning from historical operational data and maintaining certain memory states to thus analyze trends. Incorporating specially designed loss functions (such as ASUE - average safe underestimation error, or MAE - mean absolute error, used in tandem with a threshold) that penalizes overestimation of RUL further improves the model's reliability, as demonstrated in recent studies. Implementing such techniques in predictive maintenance not only optimizes maintenance schedules but also enhances the safety and reliability of aviation operations, if airlines were to employ these predictive analytics for example. Overestimation of RUL should be heavily punished, as in

realistic scenarios, there is no function beyond failure, and in the scope of aviation, these key errors in real implications can prove to be exorbitant and fatal. [2]

B. Related Works

The NASA CMAPSS Jet Engine dataset was used in a challenge competition at the International Conference on Prognostics and Health Management (PHM) in 2008, where researchers and teams competed to employ certain data analytic techniques and machine learning in order to improve prognostics for the RUL vector. Current research and attempts on the same datasets are still being continued to this day with an open-ended, no close-date, challenge. [3]

Among the top-performing approaches as of 2016, similarity-based modeling demonstrated significant effectiveness, achieving a competition score of 512.12 and a mean squared error (MSE) of 152.71. This method involved the manual selection of key sensor features—specifically sensors 7, 8, 9, 12, 16, 17, and 20—based on their continuous and consistent degradation trends (see table in methodology section on sensor allocations for further context). To construct the predictive framework, the first 5% of the data for each engine instance was labeled as the healthy state, while the remaining 95% was designated as failure data. Afterward, the data was then categorized into six bins, corresponding to six distinct operating conditions, with each bin being used to train a separate exponential regression model to characterize the progression from healthy operation to failure. The final RUL estimate was obtained by aggregating predictions from all models, and a post-processing step was applied to cap the estimates at a predefined threshold, thereby reducing the likelihood of late predictions. [4]

Recurrent Neural Networks (RNNs) have also been employed as an alternative approach, achieving a competition score of 740.31 with an MSE of 224.79 [5]. Unlike similarity-based methods, RNNs leverage functional mappings between input features and RUL to capture time-dependent degradation patterns, and to enhance predictive accuracy, a Multilayer Perceptron (MLP) classifier was initially trained to differentiate between healthy and faulty states, achieving an error rate of only 1%. However, due to the time-series nature of the data, RNNs were ultimately chosen over MLPs, as they are inherently more effective at modeling sequential dependencies and handling truncated instances. [5] The model utilized all available sensor and operational features, with gradients computed through truncated backpropagation through time, complemented by an extended Kalman filter to refine weight adjustments. To mitigate overestimation penalties, RUL predictions were capped at 130 cycles. Additionally, an evolutionary approach based on differential evolution was incorporated to improve model robustness and create an aggregate of efficient parameterization – meaning that a large number of RNNs were produced, modelled and trained, from wherein the top performing-models were selected for validation. Cross-validation on the dataset revealed that engine health degradation typically follows four distinct phases: steady operation, an inflection point or "knee," accelerated degradation, and eventual failure. [5]

Another noteworthy methodology combined MLPs with Kalman filtering techniques to enhance RUL estimation. While MLPs provided a strong functional mapping between sensor data and RUL, Kalman filters were employed to iteratively refine the model's predictions, particularly in dynamic operational conditions. This hybrid approach sought to balance computational efficiency with predictive accuracy, addressing some of the inherent limitations of purely neural network-based models. [6] The study by Ramasso and Saxena (2014) on this competition and the various methodologies employed offers a holistic analysis of the different prognostic algorithms that were applied to the C-MAPSS datasets, focusing on challenges such as sensor noise, varying operating conditions, and multiple simultaneous fault modes. By benchmarking various methods, including similarity-based models and recurrent neural networks, their study helped highlight which research teams and their methodologies' entailed certain strengths and limitations with their respective model approaches – helping guide the development of more robust predictive models. [7]

C. Problem Definition

Building upon these methodologies, this project aims to:

- address these challenges by capitalizing on the abundance of data from jet engine sensors
- develop a preliminary deep learning-based predictive model trained on NASA's C-MAPSS dataset (stemmed from the PHM08 Prognostics challenge) for aeronautic Turbofan jet engines, incorporating tree-based methods (random forests, extreme gradient boost trees) and an LSTM network (long short-term memory) to improve accuracy and minimize deviation in RUL estimation.

II. Methodology

This section outlines the approach taken to develop the predictive model for jet engine Remaining Useful Life (RUL) estimation. The process involves data preprocessing, feature engineering, model selection, training, and evaluation. All throughout development process data visualizations were appropriated to provide visual interpretation of the analysis

The following steps were taken in order to accomplish the aim of the paper.

1) **Data Preprocessing**: The NASA C-MAPSS dataset was loaded into a Pandas DataFrame, where key sensor readings and operational settings were visualized to understand their distributions. Missing values were examined and handled appropriately, ensuring data consistency before further processing. Error functions were also defined.
2) **Modelling & Proposed Solution**: normalizing the data, scaling, and labelling output vectors of actual RUL data provided the data set. For modeling, a hybrid approach was implemented using Long Short-Term Memory (LSTM) networks for capturing time-dependent

degradation patterns, along with Random Forest Regression and XGBoost to improve predictive accuracy and generalization.

3) **Evaluation & Error Analysis:** The model's performance was assessed using multiple evaluation metrics to ensure robustness and accuracy in predicting Remaining Useful Life (RUL). Root Mean Squared Error (RMSE) was used as the primary metric due to its sensitivity to large errors, making it suitable for capturing deviations in long-term degradation predictions. Additionally, Mean Absolute Error (MAE) was calculated to provide an average magnitude of prediction errors without penalizing larger deviations disproportionately. To further analyze model performance, the overall loss trend across epochs was tracked to observe how effectively the models learned from the data over time. By examining the loss curves, overfitting and underfitting were identified, guiding adjustments in model complexity and regularization techniques. These combined evaluations ensured a comprehensive understanding of the model's predictive reliability and alignment with real-world degradation patterns.
4) **Model Refinement:** Additional hyperparameter tuning was conducted for LSTM, Random Forest, and XGBoost models to optimize predictive performance. XGBoosting was later added on the initial LSTM model for model accuracy and robustness, and was also fine-tuned Further analysis included examining the impact of different sensor combinations, refining feature selection using heatmaps and correlation matrices, and assessing the significance of various preprocessing techniques through other visualizations

Important steps included acknowledging and understanding the layout and schematic of the engine, and how certain sensor data contributes to RUL data in different weightages. Understanding the influence of each sensor on the prediction is critical, as different engine parameters contribute unequally to degradation modeling. Some sensors, such as core speed (Nc) and burner fuel-air ratio (farB), have a stronger correlation with engine wear, while others may introduce noise if not properly accounted for. Identifying these varying weightages ensures a more accurate and reliable predictive model. See the engine schematic for the anatomy of a Turbofan engine.

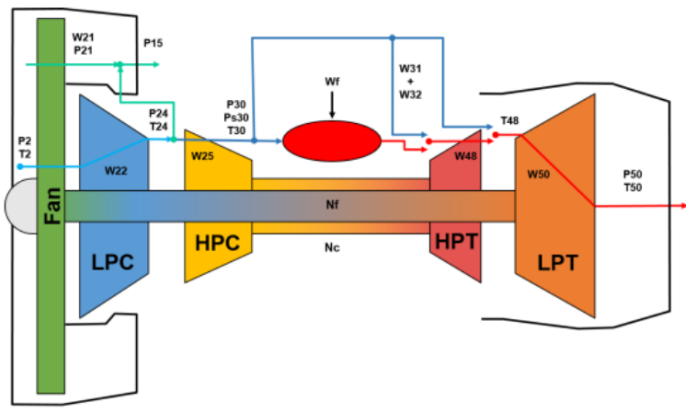

Figure 1: Schematic representation of a turbofan jet engine, illustrating key components and sensor locations used for predictive maintenance. The diagram highlights major sections, including the Fan, Low-Pressure Compressor (LPC), High-Pressure Compressor (HPC), High-Pressure Turbine (HPT), and Low-Pressure Turbine (LPT). Sensor placements for temperature, pressure, speed, and mass flow measurements are indicated, aligning with the input features used in Remaining Useful Life (RUL) prediction models.

Considering the respective sensor inputs, the following input features would accompany the finalized dataframe for modelling:

Table 1. Summary of the 26 input features used in the predictive maintenance model for jet engines. The table includes sensor number, sensor name, measured metric, and corresponding units. The features encompass operational settings, temperature, pressure, speed, fuel-air ratio, and coolant bleed measurements, all essential for modeling engine degradation and predicting Remaining Useful Life (RUL).

Sensor Numbe	Sensor Nam	Metric	Units
Sensor 1	T2	Total Temp at Fan Inlet	Rankine
Sensor 2	T24	Total Temp at LPC Outlet	Rankine
Sensor 3	T30	Total Temp at HPC Outlet	Rankine
Sensor 4	T50	Total Temp at LPC Outlet #2	Rankine
Sensor 5	P2	Pressure at Fan Inlet	psia
Sensor 6	P15	Total Pressure in Bypass-Duct	psia
Sensor 7	P30	Total Pressure at HPC Outlet	psia
Sensor 8	Nf	Physical Fan Speed	rpm
Sensor 9	Nc	Physical Core Speed	rpm
Sensor 10	epr	Engine Pressure Ratio (P50/P2)	unitless
Sensor 11	Ps30	Static Pressure at HPC Outlet	psia
Sensor 12	phi	Fuel Flow Ratio to Ps30	pps/psi
Sensor 13	NRf	Corrected Fan Speed	rpm
Sensor 14	NRc	Corrected Core Speed	rpm
Sensor 15	BPR	Bypass Ratio	unitless
Sensor 16	farB	Burner Fuel-Air Ratio	unitless
Sensor 17	htBleed	Bleed Enthalpy	unitless
Sensor 18	Nf_dmd	Demanded Fan Speed	rpm
Sensor 19	PCNfR_dmd	Demanded Corrected Fan Speed	rpm
Sensor 20	W31	HPT Coolant Bleed	lbm/s
Sensor 21	W32	LPT Coolant Bleed	lbm/s

The following data visualization demonstrates the gradual degradation of frequency and operational capability of the jet engine as the cycle (akin to a sequence and time-series metric) number increases over time. This helps visualize the gradient of degradation as the engine carries through operation:

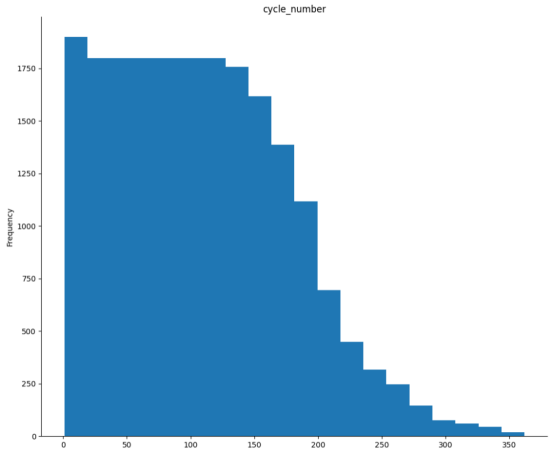

Figure 2: Histogram of cycle frequency, illustrating the gradual decline in operational cycles as engines approach failure. The decreasing frequency of higher cycle counts reflects the natural degradation process, where fewer engines remain operational at extended lifetimes.

Furthermore, on top of preprocessing, in light of feature selection — a correlation matrix is a crucial tool for understanding the relationships between different variables in a dataset. It quantifies the strength and direction of linear associations between features, helping to identify redundant variables, potential predictors, and dependencies that may impact model performance. In the context of this project, analyzing feature correlations can guide feature selection, reducing dimensionality and improving model efficiency. Strong correlations between sensor readings and Remaining Useful Life (RUL) can indicate which measurements are most predictive of engine degradation, aiding in more accurate failure forecasting, whereas weaker correlations between input features and the RUL vector can help demonstrate the weightage of each input feature respectively on the output. This aids in clarification on which features to prioritize in regression. These weights were then used to account for sensor value and impact on RUL.

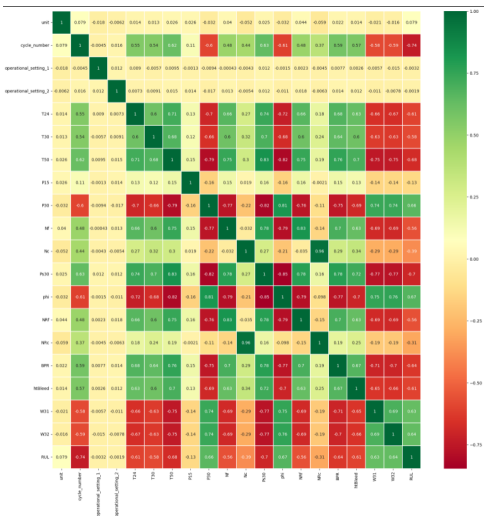

Figure 3: Heatmap visualization using Seaborn, of the correlation matrix for jet engine sensor readings and operational settings. The color intensity represents the strength of correlation between variables, with green indicating positive correlations and red indicating negative correlations. For example, T50 and P30 exhibit a strong positive correlation (0.83), while T50 and phi show a strong negative correlation (-0.82). The correlation between cycle_number and RUL (-0.79) highlights the expected relationship between engine cycles and remaining useful life. These relationships aid in feature selection.

After feature engineering, a Long Short-Term Memory (LSTM) network was set up and employed to capture the temporal dependencies in jet engine sensor data, effectively modeling the degradation patterns over operational cycles for Remaining Useful Life (RUL) prediction. The network was structured to process sequential input features, allowing it to learn patterns in engine performance decline and make time-series predictions on impending failures. Additionally, a Random Forest Regression model was utilized as a complementary approach, leveraging its ability to handle complex, nonlinear relationships between sensor readings and RUL. By aggregating multiple decision trees, this model provided robust predictions while mitigating overfitting, offering an interpretable alternative to deep learning-based methods. XGBoost was further explored as a potential enhancement, leveraging gradient-boosted decision trees to optimize predictive accuracy. In this context, XGBoost's ability to handle missing data, capture feature importance, and improve generalization makes it a strong candidate for refining RUL estimations and improving failure prognosis. For quick engagement, a user input field and aesthetic visualized tabular display were involved to allow the user to retrieve a desired quantity of predicted RUL values across the entire lifecycle (approx. 20630 entries).

III. Results

In terms of results analysis and contrasting the accuracy of RUL vector prediction, The performance of the predictive models was evaluated by comparing the predicted RUL vector against the actual RUL metric (as part of the dataset) through trendline visualizations, both before and after applying XGBoost. This comparison provided insight into the effectiveness of different modeling approaches in capturing degradation patterns. Additionally, model accuracy was assessed using three key error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and overall model loss across training epochs. RMSE was beneficial for penalizing larger errors more heavily, making it useful for detecting significant deviations in predictions. MAE provided a straightforward measure of average prediction error, ensuring interpretability. The overall model loss curve helped track convergence and assess whether the model was learning effectively over time. These metrics quantified the prediction deviations and convergence behavior, highlighting the improvements gained through boosting techniques.

The pure RUL trendline comparison between actual and predicted values prior to employing and fine -tuning XGBoosting in the forest-architecture can be observed:

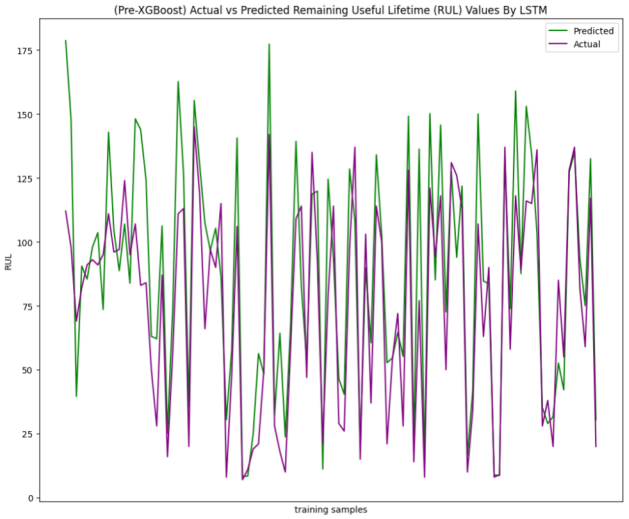

Figure 4: Pre-XGBoost Actual vs. Predicted Remaining Useful Life (RUL) Values by LSTM – The plot compares the predicted RUL (green) and actual RUL (purple) across training samples before applying XGBoost. The alignment between the two curves indicates the LSTM model's predictive capability, though noticeable deviations suggest room for improvement in accuracy and generalization.

After XGBoost optimization and hyperparameter fine-tuning, the accuracy of the model increased significantly, achieving a 87.4% model accuracy with a standard deviation of 2%.

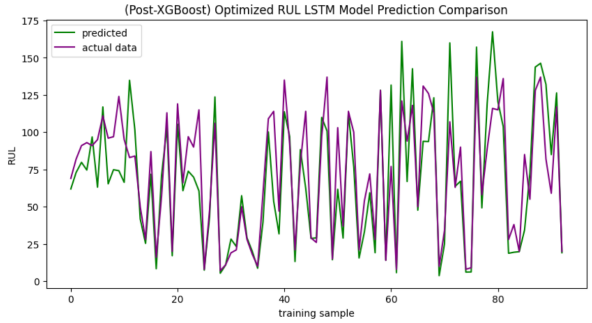

Figure 5: Figure 5: PostXGBoost Actual vs. Predicted Remaining Useful Life (RUL) Values by LSTM – The plot compares the predicted RUL (green) and actual RUL (purple) across training samples before applying XGBoost. The alignment between the two curves indicates the LSTM model's predictive capability, though noticeable deviations suggest room for improvement in accuracy and generalization

The model and error loss curves for the model AFTER XGBoosting can be observed through the training over 60 epochs:

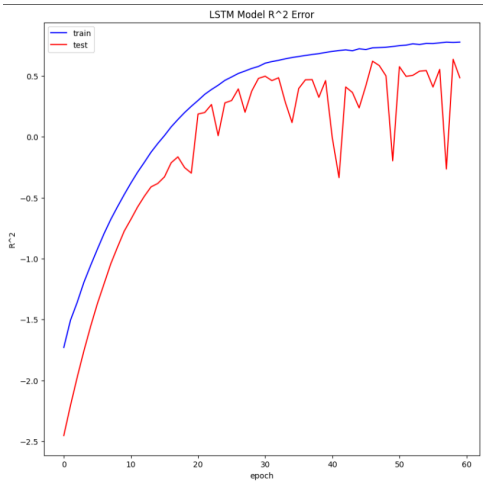

Figure 6: LSTM Model R^2 Error Curve – The plot shows the R^2 error over epochs for both training (blue) and testing (red) datasets. A higher R^2 value indicates better model performance. While the training R^2 steadily improves, the test R^2 fluctuates, suggesting potential overfitting

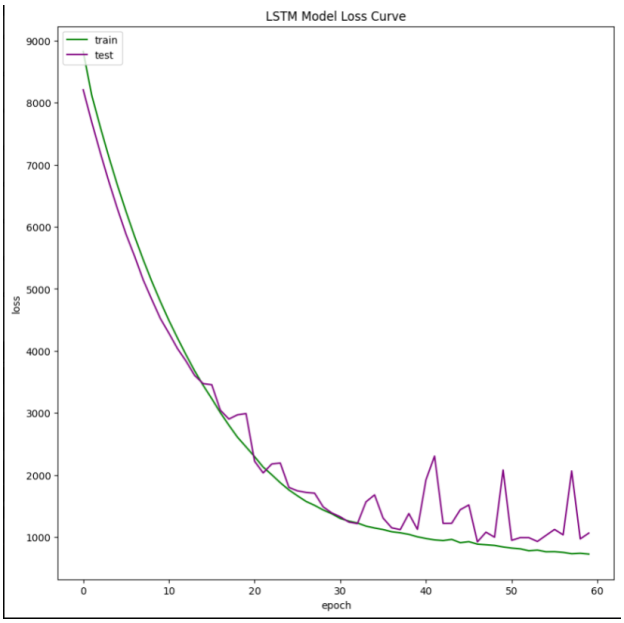

Figure 7: LSTM Model Loss Curve – The overall loss function values over training epochs for both training (green) and testing (purple) datasets. A decreasing trend suggests the model is learning effectively, but divergence between train and test loss in later epochs may indicate overfitting. Resembles MAE Curve – however overall model loss is more controlled, less overfitting, and relative overall better accuracy.

IV. Conclusion

In conclusion, this project successfully developed a preliminary deep learning model for predictive maintenance of jet engines, leveraging the NASA C-MAPSS dataset utilizing an LSTM to account for the progression and degradation of engine frequencies across cycle progression, along with data regression from feature sampling during tree-based architectures such as Random Forest regression – therein amplified with XGboost algorithms for robustness refinement. The

model was designed to effectively and relatively accurately predict Remaining Useful Life (RUL), providing a framework for early fault detection systems. Through exploratory data analysis and preprocessing, feature engineering and correlation matrix contrasting (for feature weight comparisons), and model selection, we established a robust pipeline that balances classification and regression objectives. Key insights were gained from correlation matrices, which helped assess feature importance and refine input selection. The model achieved an accuracy of 87.4% for predicted RUL output vector values, with a standard deviation of approx. 2%, indicating relative stability within the predictions – however certain data visualizations hinted at overfitting, potentially due to overly-complex modelling or intricate data. The overall procedure followed a structured approach: data preprocessing (handling missing values, scaling, and feature selection), correlation analysis to determine input significance, model training, and evaluation using appropriate performance metrics. Moving forward, enhancing model accuracy remains a primary goal. Advanced techniques such as ensemble learning, deep recurrent architectures (e.g., LSTMs or Transformers for sequential failure patterns), and hyperparameter optimization could significantly improve performance, and incorporation of differential evolution (creating batches of models and filtering for top performers) could have been beneficial. Furthermore, integrating domain adaptation strategies would enable the model to generalize across various engine types beyond the C-MAPSS dataset. Another critical next step is expanding the model's interactivity by allowing machine owners to deploy it on their custom equipment, allowing an opportunity to input custom sensor data entries, and past test sets. This would diversify the extent of deployment for models in predictive maintenance, and by facilitating user-driven data integration, the model can be retrained on specific machinery, making it more adaptable to different operational conditions. This would require developing a streamlined pipeline for data preprocessing, retraining, and deployment. Ultimately, optimizing model performance while enabling user-driven customization will be key to maximizing its practical utility in industrial applications.

REFERENCES

[1] T. Wang, J. Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for remaining useful life estimation of engineered systems," in *Proc. Int. Conf. Prognostics and Health Management (PHM)*, Denver, CO, USA, Oct. 2008, pp. 1–6.

[2] F. Heimes, "Recurrent neural networks for remaining useful life estimation," in *Proc. Int. Conf. Prognostics and Health Management (PHM)*, Denver, CO, USA, Oct. 2008, pp. 1–7.

[3] NASA, "PHM 2008 Challenge," NASA Open Data Portal, 2008, [Online]. Available: https://data.nasa.gov/Raw-Data/PHM-2008-Challenge/nk8v-ckry.

[4] ——, "CMAPSS Jet Engine Simulated Data," NASA Open Data Portal, 2018, [Online]. Available: https://data.nasa.gov/Aerospace/CMAPSS-Jet-Engine-Simulated-Data/ff5v-kuh6/about_data.

[5] A. Saxena and K. Goebel, "Turbofan engine degradation simulation data set," NASA Ames Prognostics Data Repository, 2008, [Online]. Available: https://data.nasa.gov/Aerospace/CMAPSS-Jet-Engine-Simulated-Data/ff5v-kuh6/about_data.

[6] A. Saxena, J. Celaya, E. Balaban, B. Saha, S. Saha, and K. Goebel, "Metrics for evaluating performance of prognostic techniques," in *Proc. Int. Conf. Prognostics and Health Management (PHM)*, Denver, CO, USA, Oct. 2008, pp. 1–17.

[7] E. Ramasso and A. Saxena, "Performance benchmarking and analysis of prognostic methods for cmapss datasets," HAL Archives Ouvertes, Jun. 2016, [Online]. Available: https://hal.science/hal-01324587v1/document.