Journal of
Electronics Technology Exploration

Homepage: htttps://shmpublisher.com/index.php/joetex
p-ISSN 3025-3470 | e-ISSN 3026-1066

Performance Analysis of Long Short-term Memory (LSTM) Model for Remaining Useful Life Prediction on Turbofan Engine

Themy Sabri Syuhada[1]
[1]Faculty of Mathematics and Natural Sciences Education, Universitas Pendidikan Indonesia, Indonesia

Article Info

Article history:

Received June 19, 2025
Revised June 30, 2025
Accepted July 09, 2025

Keywords:

Predictive maintenance
Remaining useful life (RUL)
Deep learning
LSTM
C-MAPSS

ABSTRACT

Accurate Remaining Useful Life (RUL) prediction is critical for the predictive maintenance and operational safety of aircraft turbofan engines. This research develops and evaluates a stacked Long Short-Term Memory (LSTM) network for RUL prediction using the NASA C-MAPSS FD001 dataset as a fundamental case study. A systematic data preprocessing pipeline was employed, including sensor selection, RUL value clipping at 130 cycles, and feature normalization to prepare the data for modeling. The LSTM model was trained with regularization techniques and an EarlyStopping callback to ensure robustness and prevent overfitting. Evaluation results on the unseen test data show the final model achieved a solid and competitive performance with a Root Mean Squared Error (RMSE) of 15.22 and a PHM08 Score of 311.20. These results demonstrate that a well-configured LSTM architecture provides a reliable baseline for engine prognostic tasks, exhibiting strong generalization capabilities on new data.

Corresponding Author:

Syuhada,
Faculty of Mathematics and Natural Sciences Education,
Universitas Pendidikan Indonesia,
Dr.Setiabudhi Street No.229 Bandung 40154, Jawa Barat, Indonesia.
Email: themysabrisyuhada@upi.edu
https://doi.org/10.52465/joetex.v3i1.585

1. INTRODUCTION

The industry 4.0 era has driven digital transformation in various sectors, where cyber-physical systems and big data analytics play a central role. In this context, Prognostics and Health Management (PHM) emerges as a crucial discipline that aims to improve the reliability, availability and safety of complex engineered systems [1]. The main goal of PHM is to move away from a schedule-based corrective or preventive maintenance paradigm, towards a more intelligent and efficient predictive maintenance strategy. By predicting when a component will fail, companies can schedule maintenance actions just in time, thereby minimizing unexpected downtime and significantly reducing operational costs [2].

One of the most critical application domains of PHM is the aviation industry, where the health of turbofan engines is a determining factor for flight safety. The ability to predict the Remaining Useful Life (RUL) of vital engine components enables proactive logistics and maintenance planning. To facilitate research in this

area without having to access classified real flight data, NASA has developed the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS), a simulation software capable of generating time series data of various engine sensors degrading to failure [3]. The dataset generated from the C-MAPSS simulation has become a standard benchmark that is widely used by the research community to develop and evaluate various RUL prediction models [4]–[6].

Along with the advancement of artificial intelligence, data-driven approaches to RUL prediction have shown very promising performance. Various studies have explored machine learning methods, ranging from conventional models to sophisticated deep learning architectures. Many researchers proposed hybrid models that combine a Convolutional Neural Network (CNN) for feature extraction from sensor data with a Long Short-Term Memory (LSTM) to model the temporal dependencies of those features [5]–[7]. Some other studies have even enhanced this architecture by adding an attention mechanism to allow the model to focus on the most relevant parts of the data when making predictions [4]. Although these complex models show highly accurate results, systematic evaluation of fundamental architectures such as LSTMs across C-MAPSS sub-datasets with different characteristics is still of great value as a solid baseline and reference point for future development of more complex methods.

Therefore, this research focuses on the development and in-depth evaluation of a Long Short-Term Memory (LSTM) model for RUL prediction using the FD001 sub-dataset as a fundamental case study. This dataset, representing a single operating condition and a single fault mode, provides a clear and crucial baseline for assessing model performance. By adopting best practices from the literature for data preprocessing, such as sensor selection and RUL value clipping, this study aims to establish a robust performance benchmark for the LSTM architecture. The paper is organized as follows: Section 2 outlines the research methodology, including the dataset description, preprocessing pipeline, and model architecture. Section 3 presents and discusses the experimental results. Finally, Section 4 concludes the findings of this study.

2. METHOD

Dataset Description

The dataset utilized in this study is the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset, developed by the National Aeronautics and Space Administration (NASA). This dataset was generated via a high-fidelity, physics-based model that simulates the degradation process of turbofan engines. Due to its public availability and realistic representation of engine degradation dynamics, the C-MAPSS dataset is widely regarded as a benchmark standard for evaluating prognostic models and has been extensively used in numerous studies on Remaining Useful Life (RUL) prediction.

The dataset is divided into four distinct sub-datasets, namely FD001, FD002, FD003, and FD004. Each sub-dataset was generated under different operational conditions and fault modes, providing a comprehensive set of scenarios for evaluating model robustness and generalization capabilities. The specific characteristics of each sub-dataset are summarized in Table 1.

Table 1. Characteristics of the C-MAPSS sub-datasets

Dataset	Train Trajectories	Test Trajectories	Operating Conditions	Fault Modes
FD001	100	100	One (Sea Level)	One (HPC Degradation)
FD002	260	259	Six	One (HPC Degradation)
FD003	100	100	One (Sea Level)	Two (HPC, Fan Degradation)
FD004	248	249	Six	Two (HPC, Fan Degradation)

While the C-MAPSS suite includes multiple complex scenarios, this study will conduct a detailed analysis focused specifically on the FD001 dataset to establish a foundational performance benchmark.

For each sub-dataset, the data are provided in space-delimited text files, which are further di-vided into training and testing sets. Each row in these files represents a single operational cycle and consists of 26 columns. The columns correspond to the unit number, time in cycles, three operational settings, and 21 sensor measurements that capture various physical properties of the engine state.

The experimental scenario simulated in the dataset follows a run-to-failure paradigm. Each time series corresponds to a unique engine, each beginning with a different, unknown degree of initial wear and manufacturing variation, which is considered normal operation. The engine operates normally at the start of each time series and develops a fault at some point during its operational life. In the training set, the fault grows in magnitude until system failure occurs. In the test set, the time series ends at an arbitrary point prior to system failure. The objective of the challenge associated with this dataset is to predict the number of remaining

operational cycles (RUL) for each engine in the test set. A separate file containing the ground truth RUL values for each test trajectory is also provided for evaluation purposes.

Data Preprocessing

A systematic data preprocessing pipeline was implemented to transform the raw time-series data into a structured and normalized format suitable for the LSTM model. This process is critical for enhancing feature relevance, ensuring model stability, and structuring the data for sequence-based learning. The pipeline consists of four sequential stages: sensor selection, RUL target transformation, data normalization, and time-series windowing.

First, a feature selection procedure was conducted to remove uninformative sensors. Based on an analysis of feature variance, sensors exhibiting constant or near-constant values throughout the engine's lifecycle were identified and excluded. These sensors provide negligible information regarding the degradation process and their inclusion could potentially introduce noise and increase model complexity without contributing to predictive performance. Specifically, sensors s_1, s_5, s_6, s_10, s_16, s_18, and s_19 were removed from the feature set.

Second, the Remaining Useful Life (RUL) target variable for the training set was generated and transformed. The ground truth RUL for each time step was initially calculated by subtracting the current cycle number from the total operational cycles of each engine. Subsequently, to better model the non-linear nature of engine degradation, where significant deterioration typically occurs closer to the end-of-life, the RUL values were clipped at an upper threshold. In accordance with the implementation in the reference notebook, a maximum RUL value of 130 was established. This technique focuses the model's learning on the most critical degradation phase, a practice widely shown to improve training stability.

Third, all input features, comprising the remaining sensor measurements and the three operational settings, were normalized to a uniform scale. The “MinMaxScaler” was employed to transform each feature into a range of [-1, 1]. It is crucial to note that the scaler was fitted only on the training data. The same fitted scaler instance was then used to transform the test data, thereby ensuring that no information from the test set leaked into the training process, which is a critical step for maintaining the integrity of the model evaluation.

Finally, the preprocessed time-series data was reshaped into a supervised learning format using a sliding window technique. This step is essential for creating input sequences and corresponding output labels for the LSTM network. A fixed sequence length of 30 times steps was utilized to construct the windows. For each engine's trajectory, the data was segmented into overlapping sequences of 30 consecutive measurements. Each sequence, a matrix of shape (30, n_features), constitutes a single input sample (X), while the RUL value at the final time step of that sequence serves as the corresponding target label (y).

Model Architecture

The prognostic model developed in this study is based on a stacked Long Short-Term Memory (LSTM) network, implemented using the Keras Sequential API to effectively capture long-term dependencies in the time-series sensor data. The architecture is composed of five sequential layers, as detailed below and illustrated in Fig. 1. The network's structure begins with an input LSTM layer comprised of 50 units, which is configured to receive sequences of 30 times steps with 17 features each. This initial layer utilizes an L2 kernel regularizer with a factor of 0.01 to penalize large weights and is set with return_sequences=True to pass the full sequence of hidden states to the subsequent layers. Immediately following this, a Dropout layer with a rate of 0.4 is applied to mitigate overfitting by randomly deactivating a fraction of neurons during training. The second stage of the network consists of another LSTM layer with a reduced capacity of 25 units, which also employs L2 regularization. As this is the final recurrent layer, its return_sequences parameter is set to False to pass only the output from the final time step onward. This is followed by an additional Dropout layer, again with a rate of 0.4, for further regularization. Finally, the architecture is concluded by a Dense output layer containing a single neuron with a linear activation function, which regresses the features from the final LSTM state into a single continuous value representing the predicted RUL.

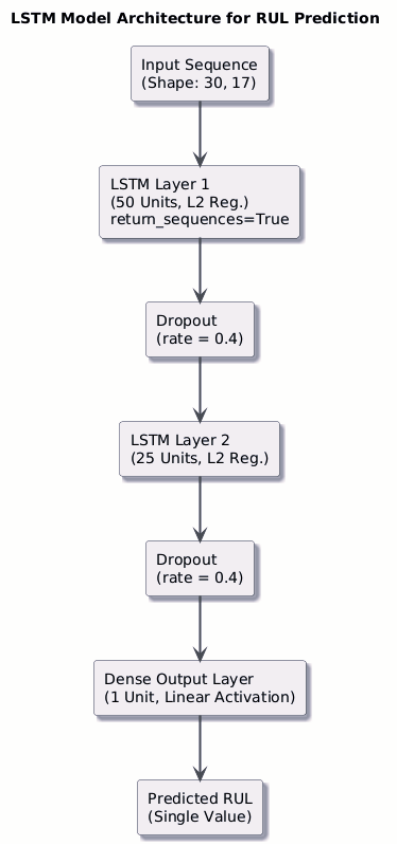

Figure 1. LSTM model architecture for RUL prediction

Evaluation Metrics

To quantitatively evaluate the predictive performance of the developed LSTM model, two distinct metrics were employed: the Root Mean Squared Error (RMSE) and the domain-specific scoring function from the PHM08 data challenge. These metrics provide complementary insights into the model's accuracy and its practical utility in a predictive maintenance context.

The primary metric for assessing the accuracy of the regression model is the Root Mean Squared Error (RMSE). RMSE measures the square root of the average of the squared differences between the predicted RUL values and the actual RUL values. It is a widely used metric for regression tasks as it penalizes larger errors more significantly and is in the same unit as the target variable, making it highly interpretable. The RMSE is calculated using the formula:

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

where n is the total number of test samples, $\hat{y}_i$ is the true RUL value for the *i*-th sample, and $\hat{y}_i$ is the RUL value predicted by the model.

In addition to RMSE, a more domain-relevant scoring function, first introduced in the PHM08 data challenge, was used for evaluation. This scoring function is designed to be asymmetric, reflecting the real-world costs associated with RUL prediction. In most predictive maintenance scenarios, predicting a failure later than it occurs (a late prediction) is significantly more detrimental than predicting it earlier than it occurs (an early prediction). Therefore, this function applies an exponentially higher penalty for late predictions. The score, S, is a cumulative sum over all test samples and is defined by the following equations:

$$\text{Let } d_i = \hat{y}_i - y_i$$

$$S = \begin{cases} \sum_{i=1}^{n}(e^{-\frac{d_i}{10}} - 1) & \text{for Early Prediction } (d_i < 0) \\ \sum_{i=1}^{n}(e^{\frac{d_i}{13}} - 1) & \text{for Late Prediction } (d_i \geq 0) \end{cases}$$

3. RESULTS AND DISCUSSIONS

This section presents the performance evaluation of the trained Long Short-Term Memory (LSTM) model. The analysis is divided into two parts: first, an examination of the model's learning behavior during the training phase, and second, a quantitative and qualitative assessment of its predictive performance on the unseen test data from the FD001 sub-dataset.

Training Performance Analysis

The model was trained for a maximum of 100 epochs, utilizing an EarlyStopping callback to prevent overfitting and a ReduceLROnPlateau callback to adjust the learning rate dynamically. The training process was automatically halted after 24 epochs, as the validation loss did not show improvement for 10 consecutive epochs. The restore_best_weights parameter ensured that the final model retained the weights from the epoch with the lowest validation loss.

The learning curves, which plot the model's Root Mean Squared Error (RMSE) on both the training and validation sets against the number of epochs, are illustrated in Fig. 2.

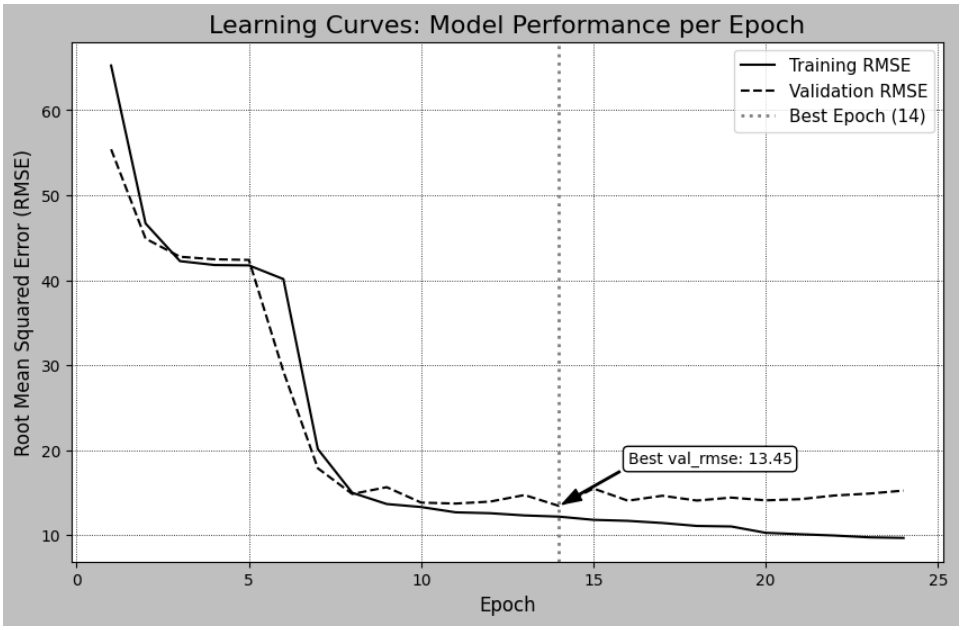

Figure 2. Training and validation RMSE curves

The learning curves reveal several key insights into the training dynamics. The model demonstrates rapid convergence in the initial phase, with both training and validation RMSE decreasing sharply until approximately epoch 14. At this point, the validation RMSE reached its minimum value of 13.45, indicating the model's optimal performance on unseen data within the training set. Beyond this point, while the training RMSE continued to decrease, the validation RMSE began to plateau and show signs of increasing, which is a classic indicator of the onset of overfitting. The “EarlyStopping” mechanism successfully identified this trend and terminated the training, preserving the model at its peak generalization performance.

Evaluation on Test Data

The best-performing model, with weights restored from epoch 14, was subsequently evaluated on the completely unseen FD001 test set, which consists of 100 unique engine trajectories. The model's predictions were compared against the ground truth RUL values to calculate the final performance metrics. The results are summarized in Table 2.

Table 2. Final performance metrics on FD001 test set

Metric	Value
Root Mean Squared Error (RMSE)	15.22
PHM08 Score (Asymmetric)	311.20

The final RMSE on the test set is 15.22, which is highly consistent with the best validation RMSE of 13.45. This small gap between validation and test performance indicates that the model generalizes well to new

data and is not significantly overfitted. The PHM08 score of 311.20 provides a domain-relevant measure of the model's effectiveness, factoring in the asymmetric cost of early versus late predictions.

To further analyze the prediction quality, a scatter plot comparing the predicted RUL against the actual RUL for all 100 test engines is presented in Figure 3.

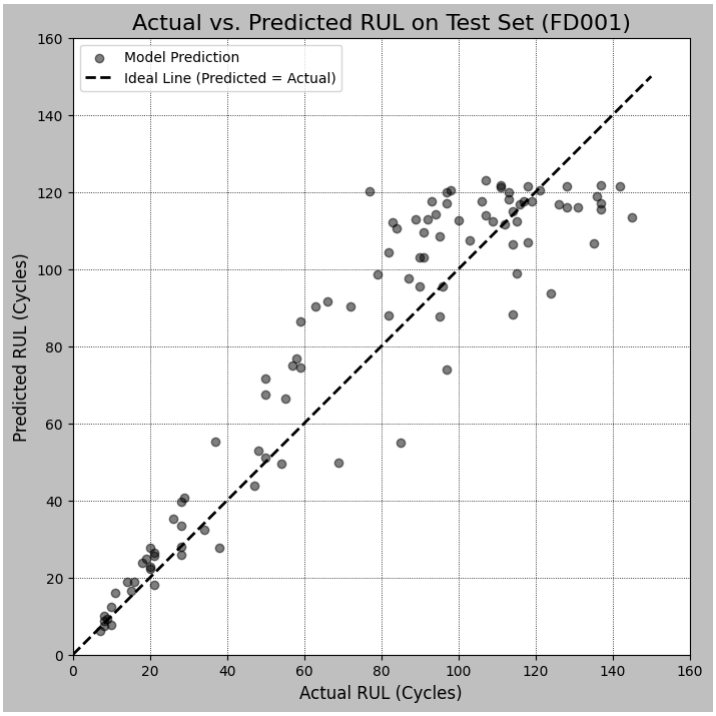

Figure 3. Actual vs. predicted RUL on FD001 test set

The plot illustrates a strong positive correlation between the predicted and actual RUL values, with most data points clustering tightly around the ideal y=x diagonal line. This visual evidence confirms the model's high predictive accuracy. The plot also reveals a slight tendency for the model to underestimate the RUL for engines that are very healthy (high actual RUL), a common characteristic for RUL prediction models, which is often mitigated by the RUL clipping technique during training.

Discussion

The experimental results demonstrate that the proposed stacked LSTM architecture, combined with a systematic preprocessing pipeline and regularization techniques, can effectively predict the RUL of turbofan engines under the FD001 scenario. An RMSE of 15.22 is a highly respectable result and is competitive with many findings in the existing literature. For instance, studies utilizing similar LSTM-based architectures have reported RMSE values in the range of 12-18 for the same dataset. While more complex hybrid models, such as those incorporating Convolutional Neural Networks (CNN) or attention mechanisms, have reported lower errors, the performance of our model establishes it as a robust and valid baseline.

The consistency between the validation and test scores is a key finding of this study, highlighting the success of the regularization strategies (Dropout, L2 regularization) and the “EarlyStopping” callback in producing a well-generalized model.

A primary limitation of this study is its focused scope on the FD001 sub-dataset, which represents the simplest scenario with a single operating condition and a single fault mode. The performance and robustness of this specific model on more complex datasets (FD002, FD003, and FD004), which involve multiple conditions and fault modes, have not been evaluated. Assessing the model's generalizability across these more challenging scenarios remains a critical direction for future work.

4. CONCLUSION

This study has successfully presented the development, training, and evaluation of a stacked Long Short-Term Memory (LSTM) network for the critical task of Remaining Useful Life (RUL) prediction in aircraft turbofan engines. By leveraging the industry-standard C-MAPSS dataset, specifically focusing on the FD001 sub-dataset as a fundamental case study, a systematic methodology was executed. This process involved a comprehensive data preprocessing pipeline, including sensor selection, RUL target value clipping, and feature normalization, followed by the implementation of an LSTM architecture with robust regularization techniques

such as Dropout and L2 regularization. The training process was carefully managed using EarlyStopping and ReduceLROnPlateau callbacks to ensure the model achieved optimal performance while effectively mitigating overfitting.

The final evaluation demonstrates the efficacy of the proposed model. On the unseen test data, the model achieved a Root Mean Squared Error (RMSE) of 15.22 and a PHM08 Score of 311.20. The strong consistency between this test performance and the validation performance (best validation RMSE of 13.45) underscores the model's ability to generalize well to new data. These findings confirm that a well-configured, fundamental LSTM architecture can serve as a highly effective and reliable baseline for RUL prognostics. The results are competitive with those reported in existing literature, establishing the validity and soundness of the implemented approach.

The primary limitation of this research is its focused scope on the FD001 dataset, which represents a single operating condition and fault mode. Therefore, the generalizability of this specific model to more complex and varied scenarios has not been assessed. Future work should extend this evaluation to the other C-MAPSS sub-datasets (FD002, FD003, and FD004) to analyze the model's robustness across multiple operating conditions and fault modes. Furthermore, exploring more advanced hybrid architectures, such as combining Convolutional Neural Networks (CNN) with LSTMs or employing attention-based models, presents a promising avenue for potentially achieving further improvements in predictive accuracy.

REFERENCES

[1] R. Moradi and K. M. Groth, "Modernizing risk assessment: A systematic integration of PRA and PHM techniques," *Reliab. Eng. \& Syst. Saf.*, vol. 204, p. 107194, 2020, doi: 10.1016/j.ress.2020.107194.

[2] M. Wu, Q. Ye, J. Mu, Z. Fu, and Y. Han, "Remaining useful life prediction via a Data-Driven Deep Learning Fusion Model-CALAP," *IEEE Access*, vol. 11, pp. 112085–112096, 2023, doi: 10.1109/access.2023.3322733.

[3] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *International Conference on Prognostics and Health Management*, 2008. doi: 10.1109/PHM.2008.4711414.

[4] C.-S. Hsu and J.-R. Jiang, "Remaining useful life estimation using long short-term memory deep learning," in *2018 IEEE International Conference on Applied System Invention (ICASI)*, 2018, pp. 58–61. doi: 10.1109/icasi.2018.8394326.

[5] L. R. Rodrigues, "Remaining useful life prediction for Multiple-Component systems based on a System-Level performance indicator," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 141–150, 2017, doi: 10.1109/tmech.2017.2713722.

[6] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Lithium-Ion battery remaining useful life prediction with Box–Cox transformation and Monte Carlo simulation," *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1585–1597, 2018, doi: 10.1109/tie.2018.2808918.

[7] D. Wang and K.-L. Tsui, "Brownian motion with adaptive drift for remaining useful life prediction: Revisited," *Mech. Syst. Signal Process.*, vol. 99, pp. 691–701, 2017, doi: 10.1016/j.ymssp.2017.07.015.