# Predictive Maintenance: A Machine Learning Approach to Remaining Useful Life (RUL) Estimation of Industrial Machines



Asmaa Motrani
Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf


Rachid Noureddine
University of Oran 2 Mohamed Ben Ahmed

# Predictive Maintenance: A Machine Learning Approach to Remaining Useful Life (RUL) Estimation of Industrial Machines

**Asmaa MOTRANI A [1], Mohamed Zakariya BOUSBAA[2], Rami Najib SAADAT[3], Rachid NOUREDDINE[4]**

[1,4] *Lab of Industrial Production and Maintenance Engineering*

[1,2,3] *Université des sciences et de la Technologie d'Oran Mohamed Boudiaf, 31036 Oran Algeria*

*E-mail: asmaa.motraniA@univ-usto.dz*

[4]*Institute of Maintenance and Industrial Safety, University of Oran2 Mohamed Ben Ahmed; P.B 1015 El M'naouer 31000 Oran Algeria .*

*E-mail: noureddine.rachid@ univ-oran2.dz*

**Abstract**: In order to create reliable models for precisely estimating the remaining usable life (RUL) of turbofan engines—a system that requires both complexity and safety—this research makes use of machine learning. To predict the RUL of industrial machines, the study focuses on building and assessing a machine-learning model. It investigates several techniques, including Random Forest, k-Nearest Neighbors (kNN), XGBoost, Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN). By utilizing these methods, the aviation sector may create sophisticated predictive maintenance solutions that maximize operational effectiveness, lower costs, and increase safety.

*Keywords*: Data-driven prognostic, Prognostic and health management, Machine learning, Turbofan engine.

## 1. INTRODUCTION

The industry has evolved towards increasingly complex systems, making breakdowns costly. Predictive maintenance has become essential to prevent failures, enabling planned interventions and reduced downtime. Prognostic and Health Management (PHM) aims to maintain systems safely and efficiently, using prognostic techniques to predict failures.

PHM can be implemented using a variety of methodologies, including history-based prognostics, data-based prognostics (Motrani, Noureddine et al. 2021; Motrani and Noureddine 2023), hybrid models, and physics-based prognostics. History-based forecasting uses expert knowledge and domain expertise to predict future behavior and equipment health, while data-based prediction uses machine learning and statistical techniques to analyze historical data and predict future behavior. Hybrid models combine knowledge-based and data-based approaches, while physics-based prognostics use physical equations and mathematical models to predict equipment behavior. The framework of our dissertation work is oriented towards the data-based prognostic approach.

Data-driven prognostics analyze sensor data using machine learning (Kimera and Nangolo 2020; Ensarioğlu, İnkaya et al. 2023) and statistical techniques to uncover patterns, trends, and relationships that indicate the health and degradation of critical components. By training predictive models on historical data, these methods can accurately predict the remaining useful life (RUL) and potential system failures. Data-driven predictions can integrate various data sources and continuously improve as new data becomes available, enabling maintenance teams to make informed decisions that optimize asset performance and maintenance activities.

Predicting remaining useful life (RUL) is a crucial aspect of predictive maintenance (Soualhi, Nguyen et al. 2024), enabling timely interventions that improve operational efficiency and mitigate catastrophic failures. This research focuses on developing robust machine learning models to accurately predict the RUL of turbofan engines, a complex and safety-critical system. Turbofan engines are the primary propulsion systems for modern commercial aircraft, and their reliable operation is paramount for aviation safety and efficiency. Accurately predicting the RUL of these engines is essential for implementing effective predictive maintenance strategies, which can optimize operations, reduce costs, and enhance safety (FAN, CHANG et al. 2023). This study leverages the power of machine learning to tackle the challenge of RUL prediction for turbofan engines. The research explores various supervised learning algorithms, including Random Forest, k-nearest Neighbors (kNN), XGBoost, Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN). The foundation of this work relies on the turbofan engine degradation simulation datasets (C-MAPSS), which effectively simulate realistic engine degradation patterns under various operational conditions. These datasets provide a realistic and challenging model development and evaluation environment, reflecting the real-world complexities inherent in turbofan engine systems. The systematic methodological framework adopted in this study encompasses crucial steps, such as data preprocessing, sensor selection, model training, and comprehensive model evaluation. The performance of the developed models is assessed using domain-specific metrics, including Root Mean Square Error (RMSE) and the PHM08 score, which account for the asymmetric costs associated with over- and underestimating RUL. The findings of this research demonstrate the effectiveness of machine learning, particularly LSTM architecture, in accurately predicting the remaining useful life of turbofan engines. The superior performance of LSTM model highlights its potential for enabling advanced predictive maintenance strategies in the aviation industry, ultimately optimizing operational efficiency, reducing costs, and improving safety. This study paves the way for future research aimed at deploying these machine learning-based solutions in real-world environments, further advancing the field of predictive maintenance and contributing to the ongoing efforts to enhance the reliability and sustainability of complex industrial systems.

The remainder of this article is structured as follows: In Section 2, a methodology framework for predicting the RUL and experimental case study is introduced. Section 3 presents a discussion of the results. Finally, conclusions and further research are addressed in section 4.

## 2. METHODOLOGY AND CASE STUDY

Predicting remaining useful life (RUL) is a cornerstone of predictive maintenance, enabling timely interventions that improve operational efficiency and mitigate catastrophic failures. This research leverages the power of machine learning to develop robust models for accurately predicting the remaining useful life of turbofan engines, a system demanding both complexity and safety. Our methodological framework adheres to a systematic and rigorous approach, comprising the following steps (Fig1). This methodology predicts the remaining life of a machine using machine learning. It begins by collecting and pre-processing data from the machine, then training a model to correlate this data with the remaining lifetime. The model is then evaluated and used to make predictions based on new data.

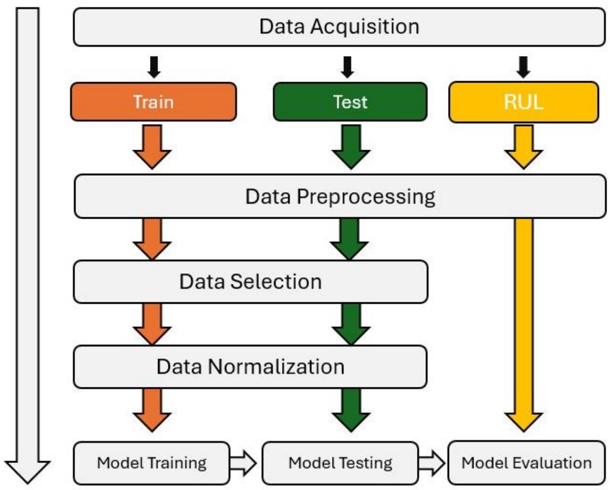



Figure 1. Methodological framework for predicting the RUL

### *2.1 Data Acquisition*

The foundation of this work relies on the turbofan engine degradation simulation datasets (C-MAPSS). These datasets effectively simulate realistic turbofan engine degradation, capturing the complex interdependencies between operational settings, sensor readings, and progression toward failure. Each dataset includes three files (Training file, Test file, and True RUL value file), each representing sensor readings of turbofan engines with the same characteristics from a single engine until their failure. The choice of the FD001 dataset is motivated by its realism, widespread use as a benchmark in predictive maintenance research, and public availability, which promotes reproducibility and encourages further research in this field.

Table 1 shows the relevant parameters of the C-MAPSS dataset and the monitoring data for each work cycle. This dataset can realistically simulate engine systems with high reliability.

**Table 1. Engine sensor data description (Saxena, Goebel et al. 2008)**

| Symbole | Description | Unité |
|---|---|---|
| T2 | Total temperature at fan inlet | °R |
| T24 | Total temperature at LPC outlet | °R |
| T30 | Total temperature at HPC outlet | °R |
| T50 | Total temperature at LPT outlet | °R |
| P2 | Pressure at fan inlet | psia |
| P15 | Total pressure in bypass-duct | psia |
| P30 | Total pressure at HPC outlet | psia |
| Nf | Physical fan speed | rpm |
| Nc | Physical core speed | rpm |
| epr | Engine pressure ratio (P50/P2) | – |
| Ps30 | Static pressure at HPC outle | psia |
| phi | Ratio of fuel flow to Ps30 | pps/psi |
| NRf | Corrected fan speed | rpm |
| NRc | Corrected core speed | rpm |
| BPR | Bypass Ratio | – |
| farB | Burner fuel-air ratio | – |
| htBleed | Bleed Enthalpy | – |
| Nf_dmd | Demanded fan speed | rpm |
| PCNfR_dmd | Demanded corrected fan speed | rpm |
| W31 | HPT coolant bleed | lbm/s |
| W32 | LPT coolant bleed | lbm/s |

### *2.2 Data Preprocessing*

Raw data often present inherent difficulties, including noise, missing values, and inconsistencies, which can negatively impact the performance of machine learning models. We have meticulously implemented the following preprocessing steps:

• **Sensor Designation**: Assigning descriptive names to each column of the datasets to improve interpretability and facilitate data manipulation.

• **Constant Value Processing**: Removing sensor data presenting constant values across all instances (Figure 2), as they offer no discriminatory power for predicting RUL.

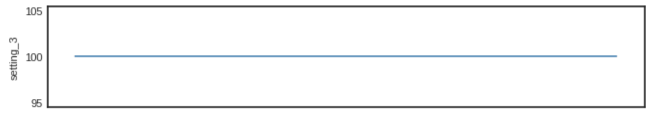



Figure 2. Setting 3 degradation

Among the data from the 21 sensors, the following sensors are removed:

- Setting 3,
- Fan inlet temperature (°R),
- Fan inlet pressure (psia),
- Engine pressure ratio (P50/P2),
- Burner fuel-air ratio,
- Demanded fan speed,
- Demanded corrected fan speed.

• **Highly Correlated Sensor Processing**: Identifying and removing sensor data from highly correlated pairs to mitigate multicollinearity, which can lead to unstable model coefficients and hinder interpretability(Asif, Haider et al. 2022).

A correlation matrix was calculated to quantify linear relationships between pairs of sensors. It presents the correlation coefficients between pairs of turbofan engine sensor data (Fig. 2). The coefficients range from -1 to 1, indicating a strong correlation (positive or negative) or a lack of correlation. Pairs (Corrected core speed and Physical core speed) showing high correlation (greater than 0.95) indicate redundancy. To mitigate multicollinearity, which can lead to unstable model coefficients and hinder interpretability, sensor data from each highly correlated pair was removed. This selection prioritized sensor data exhibiting higher variance, thus preserving more informative data.

Based on the correlation matrix, among the 14 remaining sensor data, the corrected core speed sensor was removed (Fig. 3).

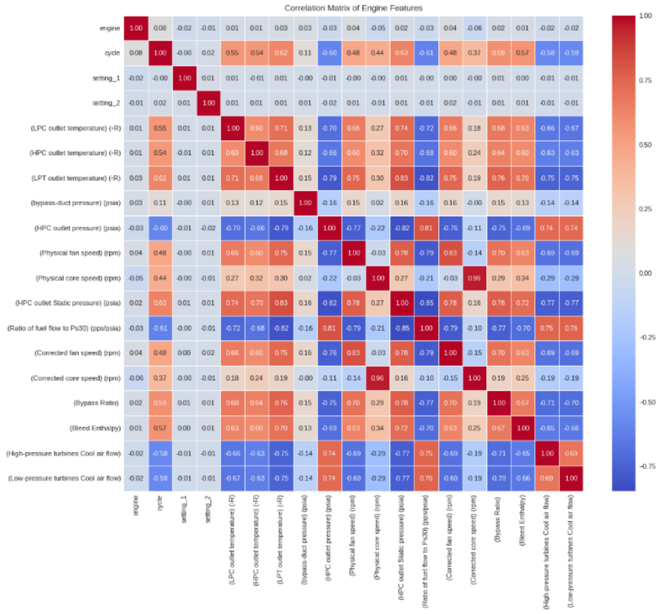



Figure 3. Correlation matrix of sensor data

*2.3 Sensor Selection*

Selecting the most informative sensors is paramount to building accurate and interpretable models. We used Lasso regression, a linear regression method with L1 regularization, for sensor selection.

The objective function of Lasso regression is as follows:

$$Minimiser : \sum(y_i - \beta_o - \sum\beta_j x_{ij})2 + \lambda\sum|\beta_j| \quad (1)$$

Where:

- $y_i$ is the target variable (RUL) for the i-th instance.

- $\beta_o$ is the intercept term.

- $\beta_j$ is the coefficient of the j-th sensor.

- $x_{ij}$ is the value of the j-th sensor for the i-th instance.

- $\lambda$ is the regularization parameter that controls the strength of the penalty.

Figure 3 illustrates the Lasso regression coefficients for each sensor data. Sensors with a non-zero coefficient were retained as RUL predictors, while those with a zero coefficient were eliminated. We observe that "(Ratio of fuel flow to Ps30)", "(High-pressure turbines Cool air flow)" and "(HPC outlet Static pressure)" are the most important sensors for RUL prediction, with relatively high positive coefficients. On the other hand, "cycle" and "(LPC outlet temperature)" show negative coefficients, indicating an inverse relationship with RUL. The remaining sensors have smaller, but non-negligible, contributions to RUL prediction.

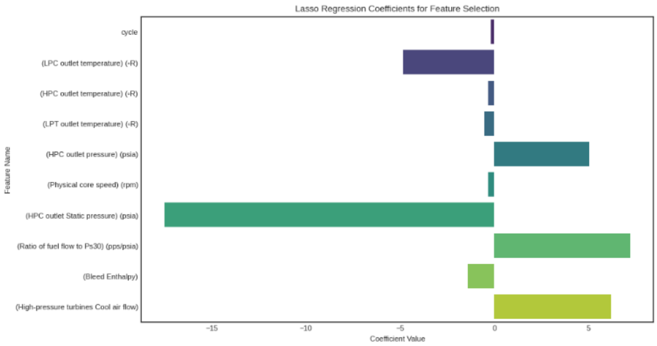



Figure 4. Lasso Regression coefficient

*2.3 Data Normalization*

Data normalization can help reduce data complexity and

Figure 6. Random Forest Structure

improve model performance and accuracy by preventing large-scale sensors from dominating the learning process. We used MinMaxScaler to transform all sensors into a common range of [0, 1], preserving the general shape of the distributions while allowing a more accurate assessment of the relative importance of each sensor. The following formula is used to normalized data :

$$Xscaled = (X - Xmin) / (Xmax - Xmin) \quad (2)$$

Where :

- $Xscaled$ is the scaled feature value.
- $X$ is the original feature value.

Figure 7. Random Forest Structure

- $Xmin$ is the minimum value of the feature.
- $Xmax$ is the maximum value of the feature.

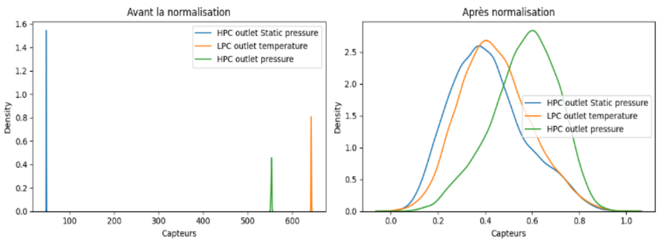



Figure 5. Sensor measurements before and after normalization

*2.4 Model Training*

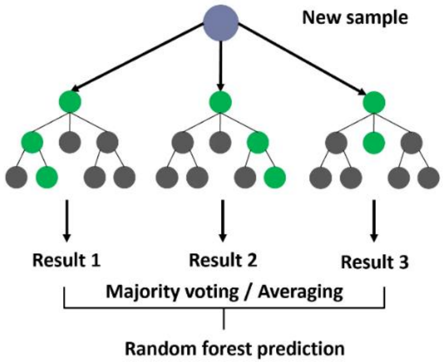
Different machine learning models were trained on the training dataset to predict the remaining useful life (RUL), including:

• **Random Forest**: An ensemble learning method that combines multiple decision trees to make predictions (Fig. 6) (Wang, Li et al. 2023).

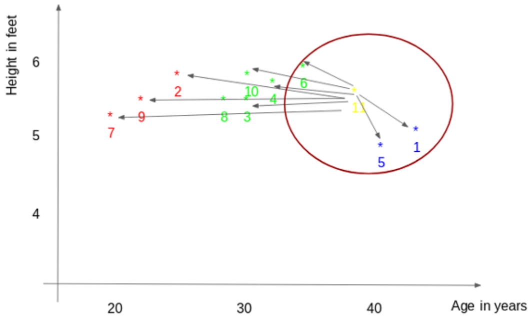
$$f_B(x) = (1/B) * \Sigma T_b(x) \quad (3)$$

Where:

- $f_B(x)$ is the final prediction of the random forest for an instance x.
- $B$ is the number of trees in the forest.
- $\Sigma T_b(x)$ represents the sum of predictions from each tree Tb for instance x.




• **k-Nearest Neighbors (kNN):** A non-parametric method that predicts based on the nearest neighbors in the feature space (Fig. 9)(Kim, Seo et al. 2021).




• **XGBoost:** A gradient-boosting algorithm known for its high performance and efficiency(Jia, Xiao et al. 2021).

$$Objective(T) = \Sigma l(y_i, y_p red, i) + \Sigma\Omega(f) \quad (4)$$

Where:

- $Objective(T)$ represents the objective function to be minimized for a given tree T.

- $\Sigma l(y_i, y_p red, i)$ is the sum of prediction errors for each instance i in the dataset,

- $\Sigma\Omega(f)$ is the sum of regularization terms for each tree f in the model, aimed at preventing overfitting.

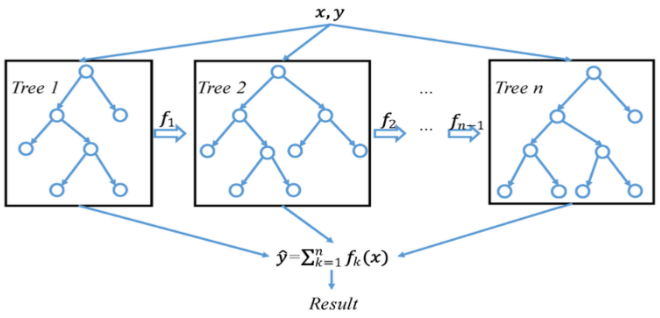



Figure 8. General structure of XGBoost

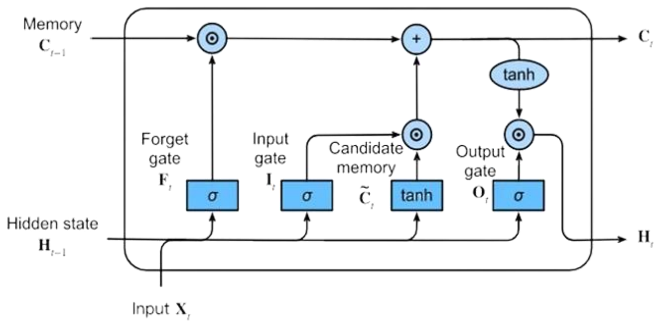



Figure 9. LSTM Structure

• **LSTM (Long Short-Term Memory):** Specialized neural networks designed to capture temporal dependencies in sequential datac(Wang, Wen et al. 2018; Asif, Haider et al. 2022).

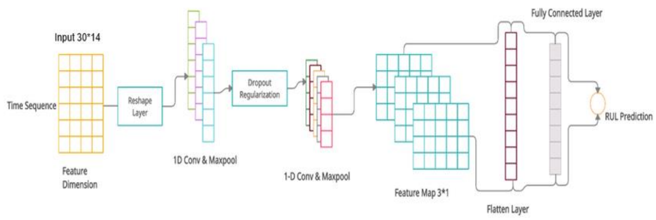



Figure 10. CNN Structure

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \quad (6)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (7)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (8)$$

$$h_t = o_t \odot \tanh(c_t) \quad (9)$$

Where:

- $i_t$ is the input gate value at time t.

- $f_t$ is the forget gate value at time t.

- $o_t$ is the output gate value at time t.

- $c_t$ is the cell state at time t.

- $h_t$ is the hidden output at time t.

- $x_t$ is the input at time t.

- $h_{t-1}$ is the hidden output from the previous step.

- $c_{t-1}$ is the cell state from the previous step.

- $W_{xi}$, $W_{xf}$, $W_{xo}$, $W_{xc}$, $W_{hi}$, $W_{hf}$, $W_{ho}$, $W_{hc}$are weight matrices.

- $b_i$, $b_f$, $b_o$, $b_c$are bias vectors.

- $\sigma$ is the sigmoid function.

- $tanh$ is the hyperbolic tangent function.

- $\odot$ represents element-wise multiplication.

• **CNN (Convolutional Neural Networks):** Deep learning models effective at extracting features from data with spatial structure, often used in image processing, but also applicable to time series data by treating them as one-dimensional signals (Fig 9)(Remadna, Terrissa et al. 2020; Khumprom, Davila-Frias et al. 2023).

*2.5 Model Testing and Evaluation*

After training, each model was applied to the independent test dataset to generate RUL predictions for each engine. The performance of the models was evaluated using the following measures:

• Root Mean Square Error (RMSE): Quantifying the average magnitude of prediction error, with lower values indicating better accuracy.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \quad (10)$$

Where $d_i$ is different between the true RUL and estimated RUL

• PHM08 Score: A domain-specific metric used in the PHM08 data challenge, specifically designed to penalize late predictions more heavily (Saxena, Goebel et al. 2008).

$$score = \sum_{i=1}^{N} S_I, \qquad S_I = \begin{cases} e^{\frac{-d_i}{13}} - 1, d_i < 0 \\ e^{\frac{d_i}{10}} - 1, d_i > 0 \end{cases} \quad (11)$$

## 3. RESULTS

Figures 11, 12, and Table 2 compare the predicted remaining useful life (RUL) values to the actual RUL values for each turbofan engine in the test set. It represents a comparison of the RMSE and PHM08 score results obtained between different models: Random Forest, KNN, XGBoost, and CNN. Based on these results, the LSTM method shows better accuracy compared to the other methods with a score of 645.07 and RMSE of 17.98.

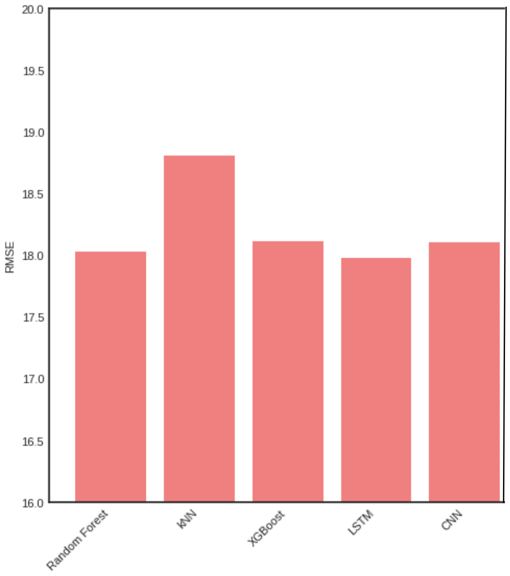



Figure 11. Comparison of RMSE result

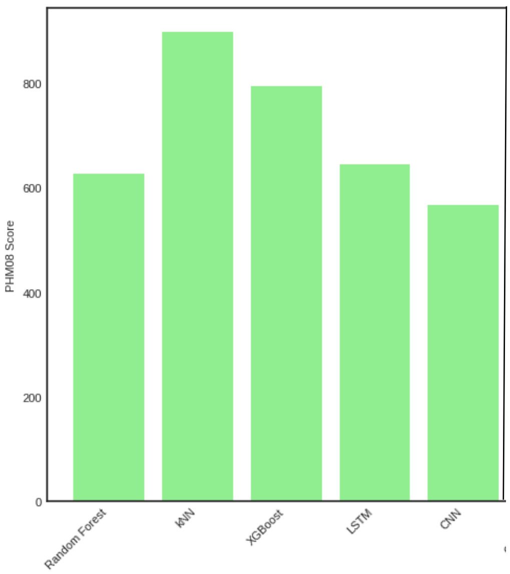



Figure 12. Comparison of PHM08 Score result

**Table 2.** Comparison of model performances

| odel | RMSE | PHM08 Score |
|---|---|---|
| Random Forest | 18.03 | 626.01 |
| kNN | 18.81 | 899.85 |
| XGBoost | 18.11 | 794.04 |
| CNN | 18.10 | 567.51 |
| **LSTM** | **17.98** | **645.07** |

## 4. CONCLUSIONS

This research demonstrates the effectiveness of machine learning approaches, particularly LSTM architectures, for accurately predicting the remaining useful life of turbofan engines. The results obtained highlight the potential of these techniques for developing advanced predictive maintenance strategies in the aviation industry. This would optimize operations, reduce costs, and improve safety. These findings pave the way for future research aimed at deploying these solutions in real-world environments.

## REFERENCES

Asif, O., S. A. Haider, et al. (2022). "A deep learning model for remaining useful life prediction of aircraft turbofan engine on C-MAPSS dataset." Ieee Access **10**: 95425-95440.

Ensarioğlu, K., T. İnkaya, et al. (2023). "Remaining Useful Life Estimation of Turbofan Engines with Deep Learning Using Change-Point Detection Based Labeling and Feature Engineering." Applied Sciences **13**(21): 11893.

FAN, Z., K.-C. CHANG, et al. (2023). "Data Fusion for Optimal Condition-Based Aircraft Fleet Maintenance With Predictive Analytics." Journal of Advances in Information Fusion **18**(2).

Jia, Z., Z. Xiao, et al. (2021). Remaining useful life prediction of equipment based on xgboost. Proceedings of the 5th International Conference on Computer Science and Application Engineering.

Khumprom, P., A. Davila-Frias, et al. (2023). A hybrid evolutionary CNN-LSTM model for prognostics of C-MAPSS aircraft dataset. 2023 Annual Reliability and Maintainability Symposium (RAMS), IEEE.

Kim, J.-T., Y.-W. Seo, et al. (2021). "A Proposal of Remaining Useful Life Prediction Model for Turbofan Engine based on k-Nearest Neighbor." Journal of the Korea Academia-Industrial cooperation Society **22**(4): 611-620.

Kimera, D. and F. N. Nangolo (2020). "Predictive maintenance for ballast pumps on ship repair yards via machine learning." Transportation Engineering **2**: 100020.

Motrani, A. and R. Noureddine (2023). "Data-driven prognostic framework for remaining useful life prediction." International Journal of Industrial and Systems Engineering **43**(2): 210-221.

Motrani, A., R. Noureddine, et al. (2021). "Performance evaluation of data-driven prognostic based on RVM-SBI technique." Journal of the Serbian Society for Computational Mechanics/Vol **15**(1): 37-50.

Remadna, I., S. L. Terrissa, et al. (2020). Leveraging the power of the combination of CNN and bi-directional LSTM networks for aircraft engine RUL estimation. 2020 Prognostics and Health Management Conference (PHM-Besançon), IEEE.

Saxena, A., K. Goebel, et al. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. 2008 international conference on prognostics and health management, IEEE.

Soualhi, M., K. T. Nguyen, et al. (2024). "Explainable RUL estimation of turbofan engines based on prognostic indicators and heterogeneous ensemble machine learning predictors." Engineering Applications of Artificial Intelligence **133**: 108186.

Wang, H., D. Li, et al. (2023). "Remaining useful life prediction of aircraft turbofan engine based on random forest feature selection and multi-layer perceptron." Applied Sciences **13**(12): 7186.

Wang, J., G. Wen, et al. (2018). Remaining useful life estimation in prognostics using deep bidirectional LSTM neural network. 2018 Prognostics and System Health Management Conference (PHM-Chongqing), IEEE.