Aircraft Engine Remaining Useful Life Prediction Using Machine Learning

Michael Kimollo, Xudong Liu
University of North Florida
John E. Mathews Jr. Computer Science
UNF Dr., Jacksonville, FL 32224
n01542707@unf.edu, xudong.liu@unf.edu

Introduction

The aviation industry faces a critical challenge: ensuring the safety and efficiency of operations while minimizing the risk of costly and potentially dangerous engine failures. Traditional maintenance approaches, often based on fixed schedules or simple usage thresholds, prove inadequate in addressing this challenge. They can lead to unscheduled downtime, unnecessary maintenance actions, and inefficient resource allocation, ultimately impacting safety, operational performance, and profitability.

Various studies have been performed to study the estimation of the Remaining Useful Life (RUL) for aircraft engines. Saxena (Saxena et al. 2008) introduced a model for damage propagation in aircraft gas turbine engine modules, providing valuable data for PHM'08. Zheng's LSTM model (Zheng et al. 2017) leveraged sequence information, demonstrating improved accuracy across PHM datasets. Lim and Fellows (Lim et al. 2014) extended the Kalman Filter ensemble with SKF, addressing non-linear degradation patterns and broadening its application. A parallel study (Li et al. 2020) introduced GRU networks for automatic feature extraction from time series data, achieving superior RUL prediction performance.

This paper delves into the development and evaluation of robust RUL prediction models for aircraft engines. It employs the NASA C-MAPSS dataset (NASA 2023), undergoes data preprocessing step, and applies a piece-wise RUL function to generate the RUL column to be predicted. By implementing and comparing various machine learning models, the project identifies the most accurate and generalizable approach for predicting RUL.

Methodology

In this study, predictive models for aircraft engine RUL involved two problem formulations, each addressing specific operational requirements and decision-making contexts. **The first is a binary classification task**, predicting whether an engine will fail within the next 30 days or not, simplifying the prediction into identifying immediate attention needs. **The second treats RUL prediction as a regression problem**, estimating the remaining operational cycles before failure, offering a detailed understanding of engine lifespan and facilitating proactive maintenance planning.

Dataset

The dataset utilized in this study is sourced from NASA (NASA 2023) and encompasses multivariate time series data from 100 unique aircraft engine units. Each unit's operational behavior is captured through readings from 21 sensors and observations of three(3) distinct operational settings. These engines, belonging to the same type, initiate their respective time series with varied levels of initial wear and manufacturing differences.

Results and Analysis

Experiment 1: Classification

In the classification experiment, the focus was solely on evaluating the efficiency of the Long Short-Term Memory (LSTM) model in predicting engine failure within one month.The optimal LSTM configuration involved two layers: the initial layer comprised 100 units, followed by a second layer with 50 units. A dropout layer with a rate of 0.2 was applied between each of these layers. The classifier achieved impressive performance metrics: 96% on precision, 0.88 recall, and 0.92 F1-score.

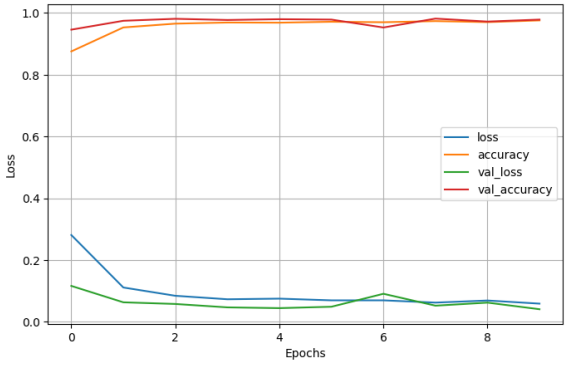

Figure 1: Training vs Validation Loss and Accuracy for LSTM Classifier

Figure 1 shows the learning curves, including the loss and accuracy of the LSTM binary classifier during training.

Experiment 2: Regression

In this experiment, three different types of models were built and tested to predict the actual RUL of the engine, and the results were compared.

Regression Models In this category, various regression machine learning models, including linear regression, random forest, k-nearest neighbors, and others, were trained to predict the RUL. Different models exhibited varying performances, as shown in Table 1. Random Forest had the best overall performance in this category, with an RMSE of 15.6 on the training set and 46.3 on the test set. However, the practicality of the random forest model was challenged by the high test set RMSE.

Table 1: Train and Test RMSE for Regression Models

Model	Train RMSE	Test RMSE
Decision Tree	0.000000	69.070572
Extra Tree	0.000000	46.190967
Forest	15.626659	46.369789
XGB	28.174315	48.496991
KNR	40.501531	48.955520
SVM Reg	43.472257	48.873759
LReg	44.660360	48.399484
Ada Reg	47.671437	51.666636

Deep Recurrent Models In this category, various deep recurrent models, including Simple Recurrent Neural Networks (RNN), LSTM, GRU, and Bidirectional LSTM, were experimented. These models were chosen due to their effectiveness with time series data, considering the time-dependent nature of sensor values. Among these models, GRU and Bidirectional LSTM demonstrated superior performance. Bidirectional LSTM showed superiority in training RMSE, while GRU outperformed in test RMSE, as summarized in the figure below:

Table 2: Performance Comparison of Deep Recurrent Models

Model	Train RMSE	Test RMSE
Bidirectional	12.292627	15.144974
GRU	12.880411	14.782531
RNN	13.034745	14.831060
LSTM	13.884804	15.078082

Deep Convolutional Models In this category, various 1D and 2D CNN underwent training, and their performances were systematically compared. Among these models, CNN-4, considered as the best CNN model, exhibited exceptional results. It consisted of two convolutional layers and two max pooling layers. This model achieved a training RMSE of 11.09 and a test RMSE of 14.02, accompanied by an R-square value of 0.88.

Additionally, other CNN models were explored, including CNN-2, which comprised 1 convolutional layer and 1 max pool layer. Furthermore, the CNN-1+LSTM model was composed of 1 convolutional layer followed by 2 LSTM layers. The corresponding performances of CNN-2 and CNN-1+LSTM are detailed in Table 3.

Table 3: Performance Comparison of selected CNN Models

Model	Train RMSE	Test RMSE
CNN-4	11.909235	14.023282
CNN-2	11.957053	16.370277
CNN-1+LSTM	12.477310	14.625450

The initial observations indicate that the proposed CNN model outperforms most of the reported values from other research studies, as indicated in Table 4.

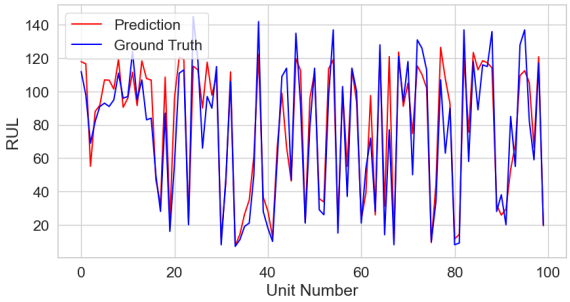

Figure 2: Final Predictions on the Test Set

Table 4: Comparison of various CNN Models in literature

Model	Test RMSE	R2	MAE
Proposed	**14.02**	**0.88**	**8.80**
Zheng et al. (2017)	16.42	NaN	NaN
Li et al. (2020)	NaN	0.78	13.45

The proposed CNN model showcased remarkable robustness in final predictions, as illustrated in Figure 2. The model's predictions exhibited a high degree of accuracy in closely tracking the actual values on the test set, thereby affirming its efficacy in predicting the RUL of the engine.

References

Li, L.; Zhao, Z.; Zhao, X.; and Lin, K.-Y. 2020. Gated recurrent unit networks for remaining useful life prediction. *IFAC-PapersOnLine* 53(2):10498–10504.

Lim, P. .; Goh, C. K.; Tan, K. C. .; and Dutta, P. . 2014. Estimation of remaining useful life based on switching kalman filter neural network ensemble. *Annual Conference of the PHM Society* 6(1).

NASA. 2023. C-mapss jet engine simulated data. `https://data.nasa.gov/Aerospace/CMAPSS-Jet-Engine-Simulated-Data/ff5v-kuh6`.

Saxena, A.; Goebel, K.; Simon, D.; and Eklund, N. 2008. Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management*, 1–9.

Zheng, S.; Ristovski, K.; Farahat, A.; and Gupta, C. 2017. Long short-term memory network for remaining useful life estimation. In *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 88–95.