Maintenance Service Events Prediction Modeling of Aircraft Gas Turbine Engines

Peeyush Pankaj[1], Shyam Joshi[2], Xiaomeng Peng[3], Reece Teramoto[4], and Taylor Hearn[5]

[1]*MathWorks India, Trillium Building, Blocks I & J, Embassy Tech Village, Bangalore, India 560103*
ppankaj@mathworks.com

[2]*MathWorks USA, 5810 Tennyson Parkway, Suite 425, Plano, TX 75024 USA*
shyamj@mathworks.com

[3,4]*MathWorks USA, 1 Apple Hill Dr, Natick, MA 01760 USA*
xpeng@mathworks.com
rteramot@mathworks.com

[5]*MathWorks USA, 5404 Wisconsin Ave, Suite 500, Chevy Chase, MD 20815 USA*
thearn@mathworks.com

ABSTRACT

This work addresses the PHM North America 2025 Conference data challenge for multi-event Remaining Useful Life (RUL) estimation on aircraft gas turbine engine modules, predicting the time-to-event for three maintenance actions: High Pressure Turbine shop visits (HPT SV), High Pressure Compressor shop visits (HPC SV), and Water Wash (WW).

We present a comprehensive workflow that integrates snapshot data quality control, virtual sensing for missing sensors (P_{25} and T_5), domain-informed feature engineering, and event-specific modeling with consensus mechanisms. Long short-term memory (LSTM) regression models are trained for HPC and WW using a custom loss function adapted from the competition, which heavily penalizes errors on early and near-term events. HPT RUL is produced by a confluence of an Artificial Neural Network (ANN) regressor and a linear degradation prior to stabilize extrapolation. A profile registration algorithm reconstructs temporal ordering in shuffled test/validation files, preserving health indicator (HI) monotonicity and degradation physics, proving a vital sanity check and building trust on the submitted results.

The MathWorks team achieved 1st place in the public test phase with the best submission score of 0.3528, proving the high quality of predictions. The functionalities and tools demonstrated in our work are generally applicable aircraft fleet maintenance services RUL predictions.

1. INTRODUCTION

Predicting the RUL of aircraft engines required maintenance service events is a long-standing challenge in prognostics and health management (PHM). Accurate prediction of maintenance events is critical not only for cost efficiency but also for safety and mission readiness. The 2025 PHM Society Data Challenge provided a dataset from the AGTF30 engine simulation, including multi-snapshot sensor measurements, flight condition data, and labels for three maintenance events of interest: HPC SV, HPT SV, and WW.

Unlike conventional RUL tasks where a single failure mode is studied, this challenge required simultaneous modeling of three interdependent service events, each with different degradation characteristics and time scales. The data also contained challenges typical of fleet operations, including missing sensors in validation/test sets, duplicates, noisy signals, and shuffled file ordering.

Our methodology is outlined in Figure 1. The key contributions of this paper are:

1. A comprehensive data cleaning and preprocessing pipeline tailored for noisy datasets under varying operating conditions.
2. Virtual sensor models to estimate missing parameters in validation/test data.

3. Domain-informed feature engineering, including pressure ratios, relative temperature drops, and efficiency proxies.
4. Design of health indicators (HIs) for compressor and turbine modules, revealing interactions between WW events and compressor degradation.
5. An ensemble of machine learning models — including ANNs, LSTMs, and survival models — tailored to each event type.
6. A novel test-time profile registration algorithm to correctly align shuffled Engine Serial Number (ESN) sequences, crucial for validating RUL predictions.

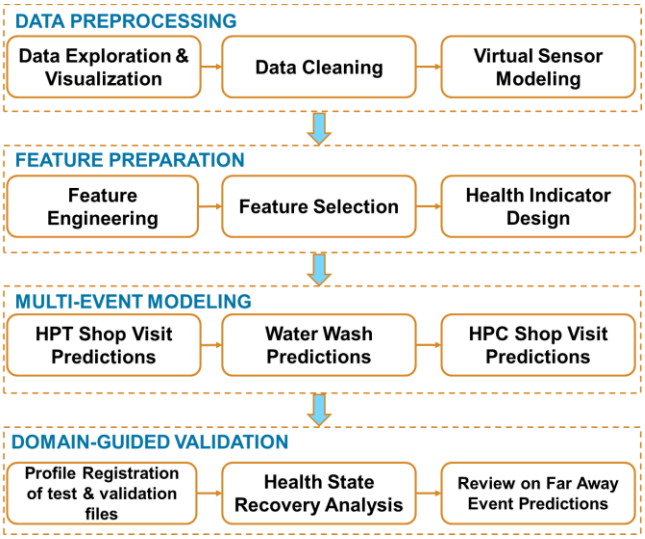


Figure 1. Workflow for multi-event RUL prediction

2. DATA PREPROCESSING

2.1 Data Exploration and Visualization

The training dataset consisted of four engine serial numbers (ESNs 101-104), each covering 20,000 operating cycles. At each cycle, engine measurements were recorded at up to eight snapshots, representing different operating conditions such as ground idle, takeoff, cruise, descent, etc. Sixteen primary sensors were provided, covering air flowpath pressures & temperatures, rotor speeds, and actuator positions. In addition, maintenance event labels were available in the form of cumulative counters and remaining cycles before the next event for HPC SV, HPT SV, and WW.

The test and validation datasets each consist of data from four separate ESNs. Specifically, the test set includes ESNs 105, 106, 111, and 112, while the validation set contains ESNs 107, 108, 113, and 114. Both datasets are made up of multiple files, with each file containing data from 150 cycles. The P_{25} and T_5 signals are not presented in both the test and validation sets. The objective of this data challenge is to predict the number of cycles remaining until the next event for HPC SV, HPT SV, and WW.

To better understand the training data and distinguish it from the test and validation datasets, we conducted initial data exploration including:

- reviewing summary statistics for each snapshot.
- analysing the flight envelope by plotting Mach number against altitude.
- visualizing sensor signals in relation to altitude.

Through this analysis, the artificially planted event markers are found in final cycles of each ESN, forcing each engine to end at 20,000 cycles as shown in Figure 2. These cycles were removed from subsequent analysis to avoid bias.

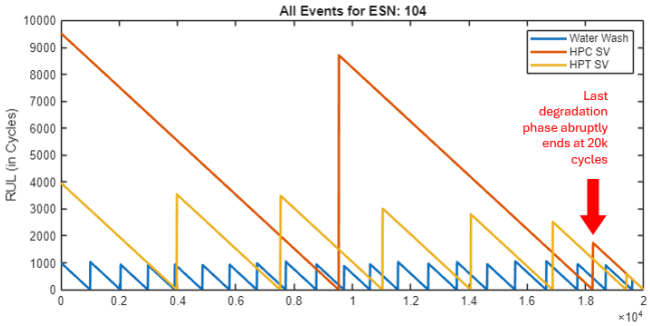


Figure 2. RUL for 3 events on the training data ESN 104

By examining the snapshot statistics, we observe that Snapshot 5 contains significantly fewer observations than the other seven snapshots as illustrated in Figure 3. This inconsistency in data recording frequency motivates us to aggregate the information from all snapshots within each cycle, thereby ensuring consistent data for every cycle. The details of this aggregation process are discussed in Section 3.1.

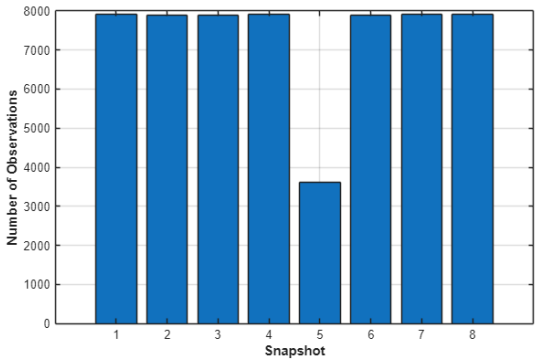
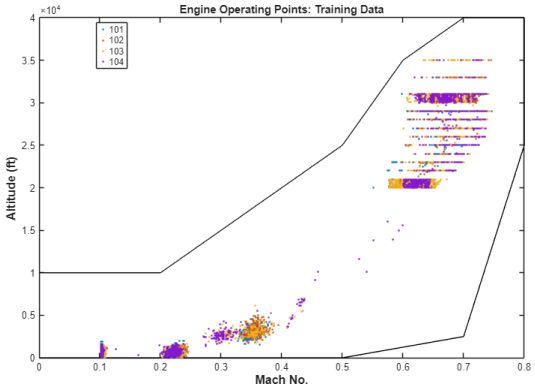


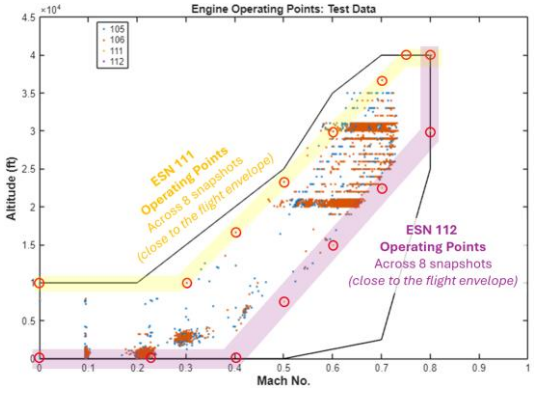
Figure 3. Training data distribution across snapshots

Operating conditions significantly affect engine performance. The Mach number represents the speed of airflow entering or exiting the engine, while altitude refers to the height above sea level. By analyzing these conditions for different ESNs, we can better understand how the units in the

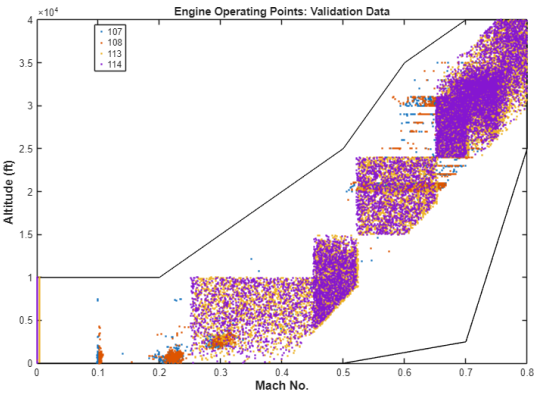
test and validation datasets differ from those in the training dataset. Figure 4 presents the flight envelope, specifically, Mach number versus altitude—for three datasets (Kratz, 2024). Certain ESNs (111 and 112 in the test set, and 113 and 114 in the validation set) exhibit a broader safe operating region, as defined by both altitude and Mach number. This broader range is associated with more rapid degradation patterns compared to other ESNs, which is highlighted in Section 5.2.



(a) Flight envelope for training dataset



(b) Flight envelope for test dataset



(c) Flight envelope for validation dataset

Figure 4. Flight envelope showing ESNs 111-114 operated closer to margins and exhibited more rapid degradation

Analysis of raw feature plots and data statistics revealed several anomalies in the dataset, including:

- duplicate rows, particularly within the metadata fields.
- negative altitude values, likely resulting from simulation artifacts.
- significant noise and inconsistent scaling across ESNs.

These issues highlight the need for normalization and outlier removal, which are addressed in the following section.

2.2 Data Cleaning

Data preprocessing was critical to ensure the robustness of downstream models' reliability. Duplicates were identified by checking for repeated values across all sixteen sensors and subsequently removed. Missing data were handled using snapshot-wise interpolation per ESN, ensuring temporal continuity.

Outlier treatment was performed on a per-snapshot, per-ESN basis. For each sensor, outliers were defined as elements more than 1.5 interquartile ranges above and below the upper and lower quartiles, respectively. Outliers were removed using MATLAB's "rmoutliers" function. Figure 5 shows the data range before and after cleaning. This strategy ensured that spurious spikes did not propagate into HIs or learned models.

After cleaning, the dataset retained sufficient coverage across all eight snapshots for meaningful feature extraction.

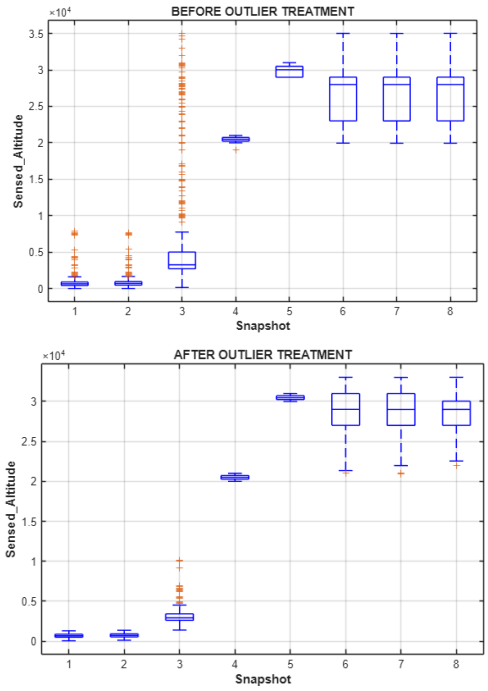


Figure 5. Boxplots of sample Altitude sensor before (top) and after (bottom) outlier treatment across snapshots in the training

2.3 Virtual Sensor Modeling

A unique aspect of this year's challenge was the absence of two key sensors, P_{25} and T_5, in the validation and test sets. To overcome this, we developed linear regression-based virtual sensor models.

Inputs included upstream and downstream sensors, snapshot labels and ambient conditions. Multiple machine learning models were trained and compared using the $R^2, RMSE, MAE$ and $MAPE$ metrics. The Interaction Linear Regression model was selected based on performance on randomly split training (80%) and test data (20%) from the competition's training dataset. The RMSE for the P_{25} and T_5 virtual sensors on the partitioned test data was 0.039 and 0.781 respectively, as shown in Figure 6. These virtual sensors were subsequently used to populate missing values in the validation and test sets, ensuring consistent feature sets across all phases of the competition.

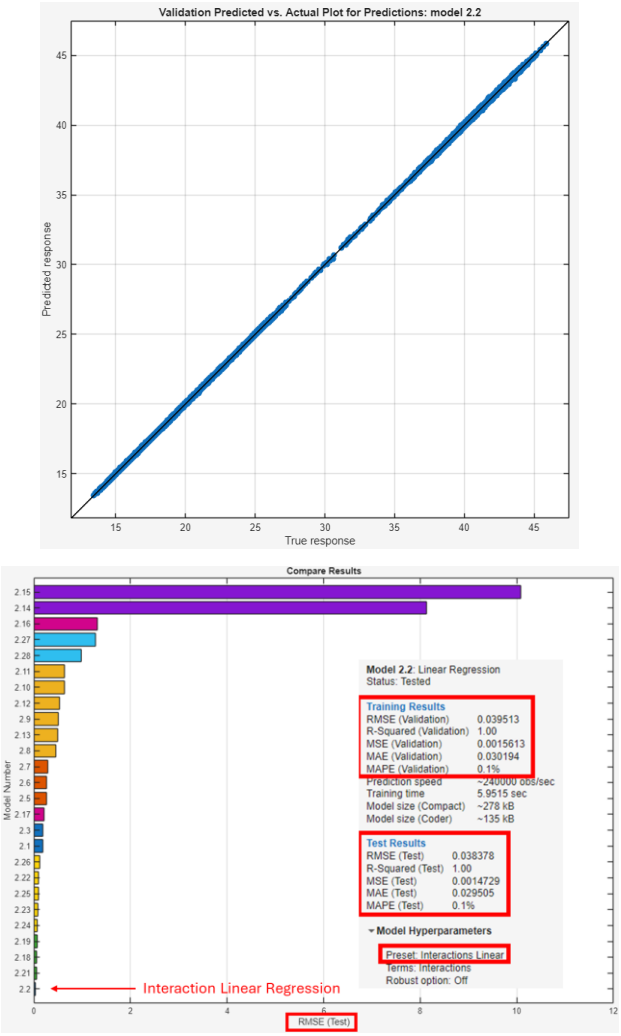


Figure 6. Interaction Linear Regression model was selected for P_{25} virtual sensor amongst several models. The same approach was used for the T_5 virtual sensor model.

3. FEATURE PREPARATION

3.1 Feature Engineering and Selection

Feature engineering and selection are critical steps in the engine analysis process, as they directly influence the performance and interpretability of predictive models. Extracting relevant features from raw sensor data enables a deeper physical understanding the engine system.

The dataset comprises a comprehensive set of engine sensor signals, including measurements such as altitude, Mach number, ambient and total pressures (P_{amb}, Pt_2), total air temperature (TAT), fuel flow (W_{Fuel}), variable area fan nozzle (VAFN), variable bleed valve (VBV), and key rotational speeds (Fan_Speed, Core_Speed). Additionally, it captures critical thermodynamic states at various engine stations, such as temperatures (T_{25}, T_3, T_{45}, T_5) and pressures (P_{25}, Ps_3). To better capture the degradation trends, a series of domain-informed features were extracted to characterize engine performance and health:

1. Pressure ratios across fan, HPC, LPC, and the overall compressor system.
   - HPC pressure ratio $= Ps_3/P_{25}$
   - LPC pressure ratio $= P_{25}/Pt_2$
   - Fan pressure ratio $= Pt_2/P_{amb}$
   - Compressor pressure ratio $= Ps_3/Pt_2$
   - Overall engine pressure ratio $= Ps_3/P_{amb}$
2. Relative temperature drops across the HPT, LPT, and combined turbine modules with respect to the corresponding entry temperature.
   - HPT relative temp drops $= \frac{T_{45} - T_3}{T_3}$
   - LPT relative temp drops $= \frac{T_5 - T_{45}}{T_{45}}$
   - Turbine relative temp drops $= \frac{T_5 - T_3}{T_3}$
3. Proxies for thermal efficiency and fuel efficiency derived from enthalpy balance approximations.
   - Thermal efficiency proxy $= \frac{T_5 - TAT}{T_3 - TAT}$
   - Combustor efficiency proxy $= \frac{T_{45} - T_{25}}{T_{25} - TAT}$
   - Compressor Thermal efficiency proxy $=$
     $$1 - \frac{1}{\text{Overall pressure ratio}^{\frac{\gamma-1}{\gamma}}}$$
   - Specific Fuel consumption proxy $= \frac{W_{Fuel}}{T_5 - TAT}$
   - Specific power proxy $= T_5 \times Mach$

4. Corrected rotor speeds using TAT.
   - Corrected fan speed $= \frac{Fan\_Speed}{\sqrt{TAT}}$
   - Corrected core speed $= \frac{Core\_Speed}{\sqrt{TAT}}$

These domain knowledge-based features enable robust monitoring, diagnostics, and performance analysis of the engine system.

As noted in Section 2, some snapshots were missing in the raw data. To address this, we summarized the snapshot data for each cycle using statistical measures such as mean, standard deviation, minimum, maximum, range, median, and RMS. This approach transforms the original eight-snapshot data into a compact set of cycle-level features.

To further reduce the dimensionality of the features, we selected key features based on the feature variance, and dropped the features with variance lower than 0.01. Those selected features were utilized in HI design and modelling process.

3.2 HEALTH INDICATOR DESIGN AND FEATURE FOR WW EVENT

To capture progressive degradation, we designed HIs for the HPC and HPT modules using MATLAB's Health Indicator Designer as shown in Figure 7. These indicators were developed using cycle-level statistical features, which were normalized to range from 1 (indicating a healthy state) to 0 (indicating failure). The health index is calculated as a weighted sum of these features, with each feature assigned a specific weight reflecting its contribution to the overall index (Zou, Hui & Hastie, 2005; Moradi, Morteza, et al, 2023).

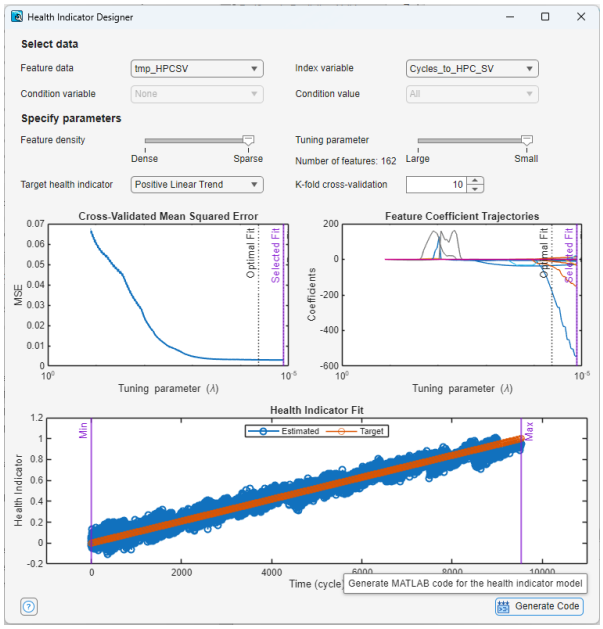


Figure 7. Health Indicator Designer result for HPC SV event

Interestingly, the HPC HIs showed a step-change in behaviour after each WW event. As illustrated in Figure 8, each WW partially restored the indicator, thereby delaying the onset of compressor degradation. This observed pattern accurately represents the real-world effect of water washing in extending the life of the HPC module, which we incorporated into our WW and HPC modeling strategies. The details of the feature are described in the Section 4.2.

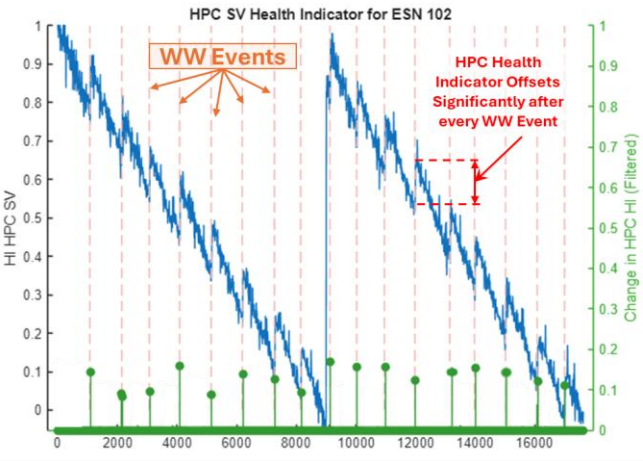


Figure 8. Example HI (left axis, blue) for HPC Degradation for ESN 102 with WW events (vertical dashed lines) and stem plot of change in HI (right axis, green)

4. MULTI-EVENT RUL MODELING

In this section, we present separate modeling strategies, each tailored to each event type.

4.1 HPT Shop Visit Predictions

For the HPT SV prediction, we trained several machine learning models and ANNs on the cycle-level feature set. Predictions were compared among different families of models on the randomly split training and test sets with 5-fold cross-validation, and an ANN model was observed to perform best among several machine learning algorithms as shown in Figure 9. The ANN was later fine-tuned to improve accuracy even further.

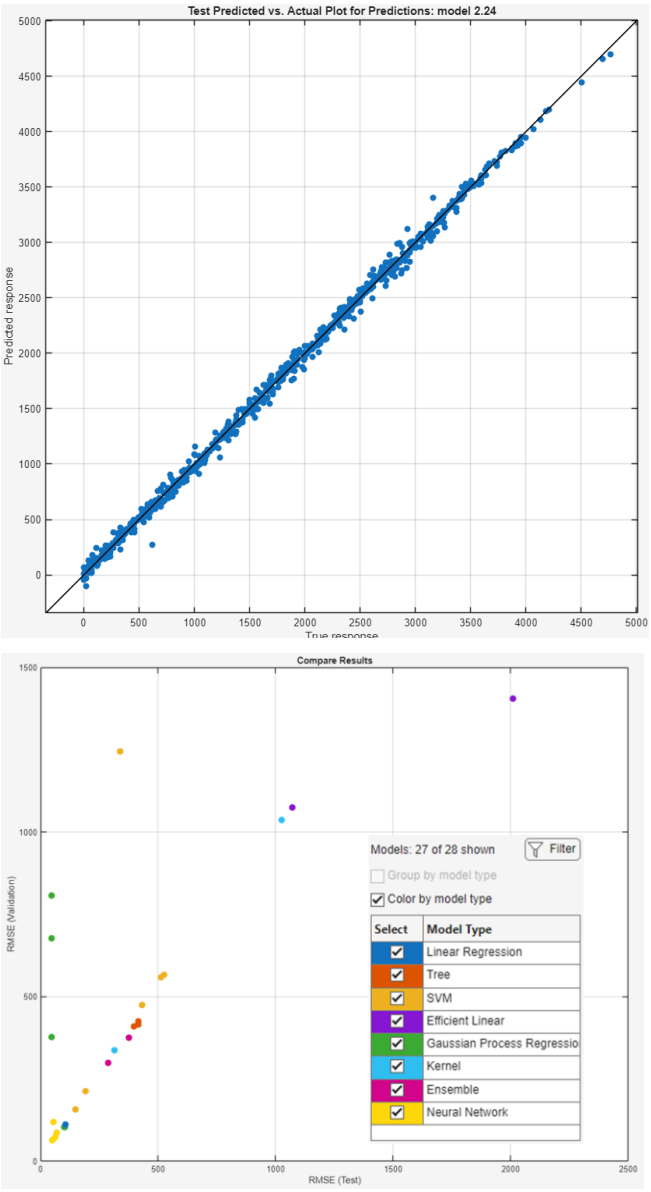


Figure 9. Model Comparison for HPT SV Predictions. ANNs performed best on the training data

4.2 Water Wash Events Prediction

An LSTM sequence-to-sequence regression model was trained on time-series features (Saxena, Goebel, Simon & Eklund, 2008). To optimize the model, we used the competition's time-weighted scoring function as a custom loss function during training.

For training, we used the cycle-level features from ESNs 101-103. ESN 104 was reserved for validation and testing. Since each file in the validation and test sets contains 1500 cycles, we also divided the training data into multiple sub-sequences, each with 1,500 cycles. From ESN 104, 20% of these sub-sequences were used for validation, and the remaining 80% were used for testing.

As described in Section 3.2, we observed that HPC HIs tended to drop by a consistent amount prior to WW events and recovered significantly after the WW. This pattern provided another cue for prediction. Hence, a binary feature was included in the training dataset to have a value of 1 wherever the HPC HI recovery is observed, and 0 for the rest of the observations.

Incorporating this WW specific feature and custom loss function, the model performed well on both validation and test part in the training dataset, as shown in Figure 10.

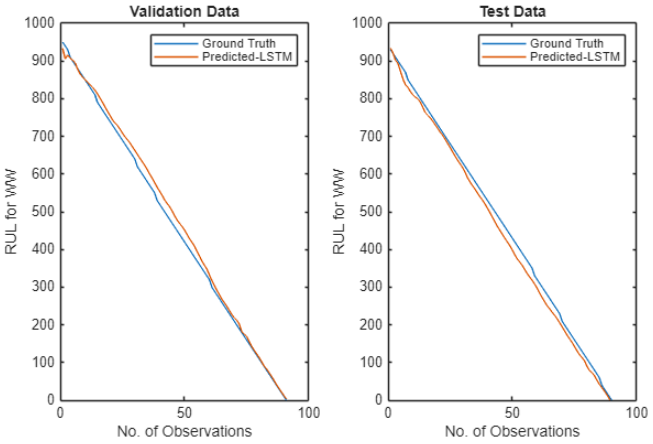


Figure 10. WW LSTM model predictions from training set on ESN 104

4.3 HPC Shop Visit Predictions

For the compressor, we trained an LSTM model with a custom time-weighted loss function, identical to that used in the competition scoring (Hochreiter and Schmidhuber, 1997). This loss penalized late predictions relative to the true RUL more than early predictions, especially near event cycles as shown in Figure 11. Both engineered features and HIs were used as inputs, along with binary labels for recovery in the HI, making this the most comprehensive of our models.

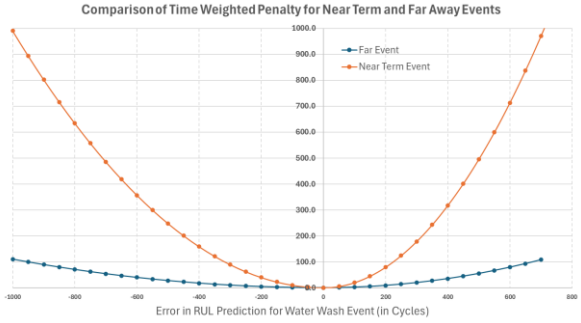


Figure 11. Time-weighted Error Function for the WW and HPC event derived from the training data, demonstrates that the competition's scoring function significantly penalizes the late predictions made for near-term events as compared to far events

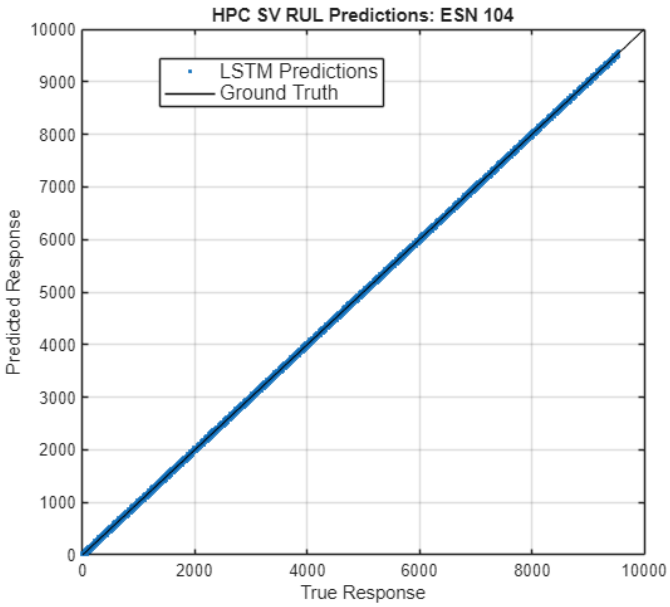


Figure 12. Ground Truth vs LSTM model predictions for HPC SV

5. Verification & Validation Of Predicted Results5.1 Profile Registration of Test and Validation Files

All ESNs in the training data were stopped at 20,000 cycles, causing the final HPC, HPT, and WW event predictions to end abruptly rather than exhibiting the typical degradation trend. This presented a significant challenge, as the test and validation datasets also featured a randomized ordering of files for each ESN. Such shuffling disrupted the temporal sequence necessary for sequence modeling and maintaining health index coherence.

To address these issues, we incorporated domain knowledge and observational insights to reorder the files appropriately. As part of our verification and validation process, we developed an optimization-based profile registration algorithm to reconstruct the correct temporal sequence. The algorithm evaluates the HI in both HPC and HPT events, minimizing score discrepancies between consecutive files. By aligning HI trajectories and degradation trends, the algorithm effectively "stitches" the files into their true chronological order.

This process is crucial for ensuring accurate RUL predictions, particularly for the final events, and was a key factor in achieving a top ranking during the test phase. Given the abrupt stops at 20,000 cycles, the final HPC, HPT, and WW events often deviate from typical degradation patterns. To ensure precise predictions for these last events, a manual prediction adjustment was applied based on the results of the profile registration. This step was essential for delivering accurate outcomes for the final HPC, HPT, and WW events.

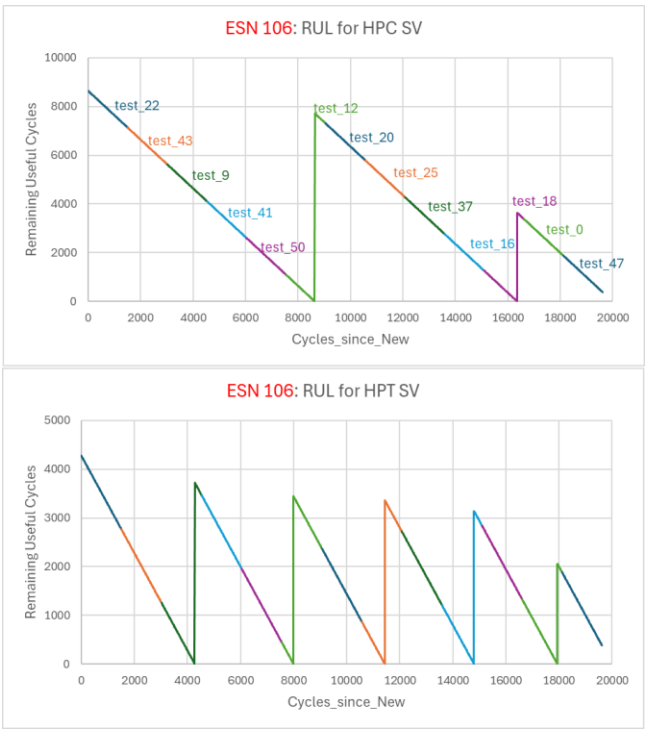


Figure 13. Model predictions shown for one of the ESNs in test data (ESN 106) after profile registration. It reflects the proper temporal order for both HPC & HPT degradation

5.2 Health State Recovery Analysis

As part of the validation and verification process, we analyzed the temporal sequence of HI profiles for each test and validation file following the profile registration step. One key observation from the training dataset—also confirmed by domain experts—is that the recovery in health state after major HPC and HPT service events is progressively reduced compared to earlier service events. This indicates that the health state continues to degrade even after each HPT or HPC shop visit.

Additionally, as discussed in Section 2.1, ESNs 111 and 112 in the test data operate closer to the flight envelope compared to ESNs 101-106 from the training and test datasets. The effects of such operational conditions are evident, as these engines exhibit more frequent HPC and HPT service events, as illustrated in Figure 14.

These observations confirm that our predictions are consistent with established physical understanding and accurately reflect the expected degradation behavior.

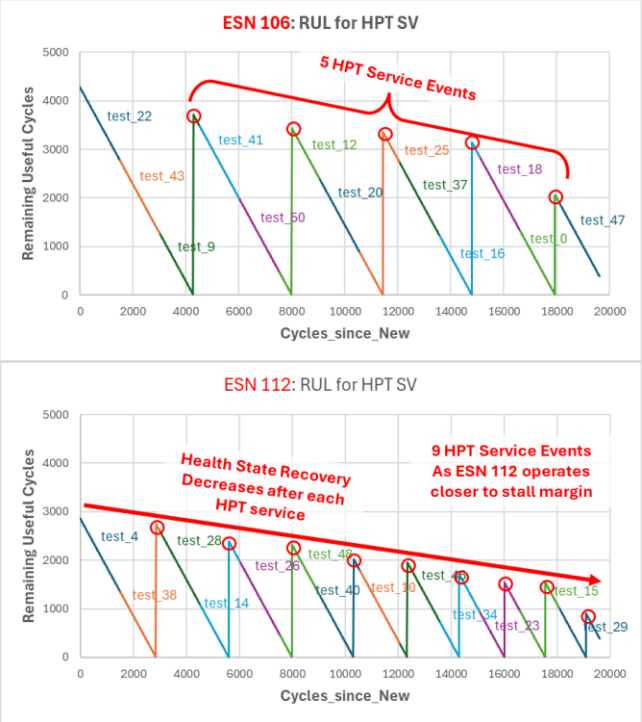


Figure 14. RUL prediction from model reflects lower health state recovery after each service event compared to the previous. More service events for ESN 112 compared to ESN 106 in the test data.

5.3. Review of Far Away WW Event Predictions

While the profile registration validation check was a breakthrough, we understood the impact of the competition's scoring function on late predictions made by the AI models. Thus, the submitted predictions on the test and validation files were carefully examined to ensure that the far away predictions made for the water wash events are backed by clear evidence of HPC HIs offset before the end of file as shown in Figure 15 for ESN 106 in test data.

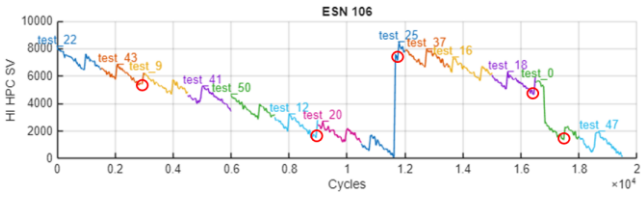


Figure 15. Careful Examination of Offset in HPC HI slope when far away WW predictions are submitted at end of these files

These sanity checks helped us achieve strong results in the test phase, and we also took a conservative approach on some of the WW predictions to avoid large penalties on the validation data, rather than chasing a perfect score.

6. CONCLUSION

We presented a comprehensive workflow for multi-event RUL prediction in gas turbine engines, as part of the 2025 PHM Society Data Challenge. Our approach combined careful data preprocessing, virtual sensor modeling, domain-informed feature engineering, HI design, and ensemble learning methods.

The methodology successfully captured degradation patterns for HPC, HPT, and WW events, achieving 1st place in the test phase of the competition.

7. ACKNOWLEDGEMENT

The authors acknowledge support and encouragement from MathWorks for participating in PHM Data Challenge 2025. Extensive technical computing resources, software tools, IT infrastructure, and the collaborative support of fellow MathWorkers are thankfully acknowledged.

8. REFERENCES

Kratz, Jonathan L. (2024) "The Advanced Geared Turbofan 30,000 lbf – electrified (AGTF30-e): A Virtual Testbed for Electrified Aircraft Propulsion Research", AIAA 2024-3824. https://doi.org/10.2514/6.2024-3824

Moradi, Morteza, et al. "Intelligent Health Indicator Construction for Prognostics of Composite Structures Utilizing a Semi-Supervised Deep Neural Network and SHM Data." Engineering Applications of Artificial Intelligence, vol. 117, Jan. 2023, p. 105502. DOI.org (Crossref), https://doi.org/10.1016/j.engappai.2022.105502.

Zou, Hui, and Trevor Hastie. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, Apr. 2005, pp. 301–20.

Saxena, Abhinav, Kai Goebel, Don Simon, Neil Eklund. " Damage propagation modeling for aircraft engine run-to-failure simulation." In Prognostics and Health Management, 2008. PHM 2008. International Conference on, pp. 1-9. IEEE, 2008.

Hochreiter, S., and J. Schmidhuber. "Long short-term memory." *Neural computation*. Vol. 9, Number 8, 1997, pp.1735–1780.