

Multi-Sensor Kalman Filter Fusion for Remaining Useful Life Prediction of Commercial Jet Engines

Ondřej Baštař

Czech Technical University in Prague, Faculty of Electrical Engineering

bastaond@fel.cvut.cz

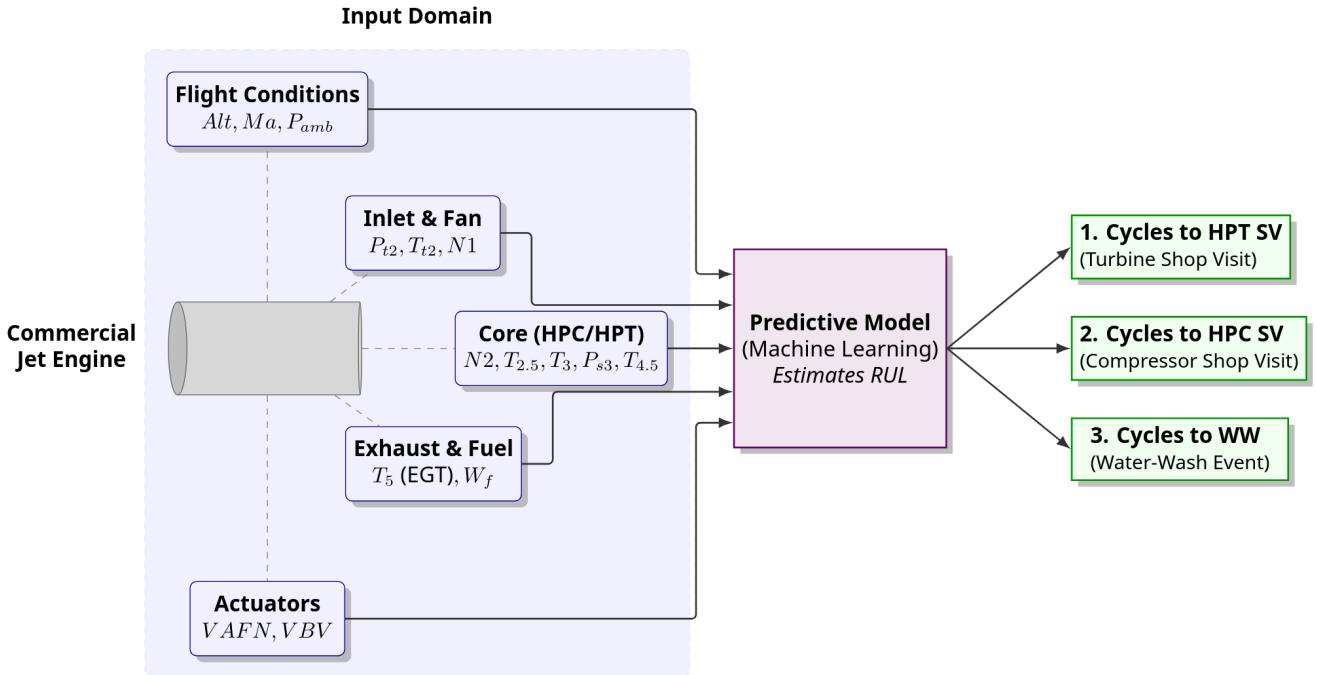


Fig. 1. Diagram describing the PHM Society North America 2025 Data Challenge

Abstract—This report presents a solution for the PHM Society North America 2025 Data Challenge, which requires estimating the Remaining Useful Life (RUL) for three distinct maintenance events: Water Wash (WW), High-Pressure Compressor (HPC) Shop Visit, and High-Pressure Turbine (HPT) Shop Visit. The proposed approach utilizes a hybrid methodology combining physics-inspired signal processing with machine learning. First, a baseline model normalizes sensor data against operating conditions. Second, a bank of Kalman Filters estimates the underlying health state and degradation rate (slope) of key engine sensors. Finally, these estimated states are used as features for a Random Forest regressor, followed by a secondary smoothing Kalman Filter to stabilize RUL predictions. The method achieves a weighted score of 37.31, significantly outperforming a baseline Gradient Boosting approach (score 93.60).

Index Terms—Predictive Maintenance, Kalman Filter, Remaining Useful Life, Machine Learning, Turbofan Engine

I. INTRODUCTION

The reliability of commercial jet engines is paramount for the aviation industry. Modern Engine Health Management (EHM) systems rely on sensor data to predict the Remaining

Useful Life (RUL) of critical components, allowing for timely maintenance and reducing operational costs.

The objective of this work is to address the challenge posed by the PHM Society North America 2025 Data Challenge [1]. The task involves predicting the time, in flight cycles, until three specific events occur:

- 1) **HPC Water Wash (WW):** A routine maintenance action to clean the compressor.
- 2) **HPC Shop Visit:** Major maintenance for the High-Pressure Compressor.
- 3) **HPT Shop Visit:** Major maintenance for the High-Pressure Turbine.

The dataset consists of snapshot sensor data (temperatures, pressures, speeds) collected during various flight phases. A key challenge is the asymmetric and time-weighted scoring metric, which heavily penalizes late predictions that could lead to in-service failures.

The provided dataset presents a significant challenge for data-hungry algorithms. It contains measurements from only 4 engines, with approximately 2000 data points per engine across various flight states. Crucially, the target events are

sparse: each engine averages only 20 Water Wash events, 3 HPC Shop Visits, and 6 HPT Shop Visits. This scarcity of failure data (3 samples for HPC visits) necessitates a method that relies on signal processing and domain knowledge rather than purely statistical inference.

II. RELATED WORK

State-of-the-art approaches to RUL estimation generally fall into three categories: physics-based, data-driven, and hybrid approaches.

Physics-based models rely on detailed mathematical models of degradation (e.g., crack propagation), offering high interpretability but requiring precise system parameters often unavailable in public datasets [2].

Data-driven approaches, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, directly map raw sensor data to RUL [3]. While powerful, they are often susceptible to sensor noise and struggle with generalizing across different operating conditions without massive datasets.

Hybrid approaches attempt to bridge this gap. A common technique involves using filtering methods, such as Kalman Filters or Particle Filters, to extract a clean “health index” from noisy data, which is then projected forward [4]. This work adopts a hybrid strategy...

III. METHODOLOGY

The proposed solution processes the raw engine data through a multi-stage pipeline: normalization, state estimation, regression, and smoothing. Overview of the pipeline is shown in Fig. 2.

A. Data Preprocessing and Normalization

The raw dataset includes 16 sensor features, comprising physical temperatures ($T_{t_2}, T_{2.5}, T_3, T_{4.5}, T_5$), pressures ($P_{t_2}, P_{s_3}, P_{2.5}$), rotor speeds (N_1, N_2), and actuator states.

Raw sensor measurements z_{meas} are highly dependent on flight conditions (altitude, mach, ambient pressure). To isolate component degradation, I first train a “healthy baseline” model using second degree polynomial regression on the initial 100 cycles of the engine’s life. This was chosen

as linear regression cannot capture the nonlinear physical behavior.

For a given sensor S , the expected value \hat{z} is predicted based on environmental features E . E is composed of flight conditions (altitude, mach, ambient pressure, fuel weight, ambient temperature):

$$\hat{z}_k = f_{\text{baseline}}(E_k). \quad (1)$$

The residual r_k , representing the deviation from healthy behavior, is calculated as:

$$r_k = z_{\text{meas},k} - \hat{z}_k. \quad (2)$$

This residual serves as the primary input for health tracking.

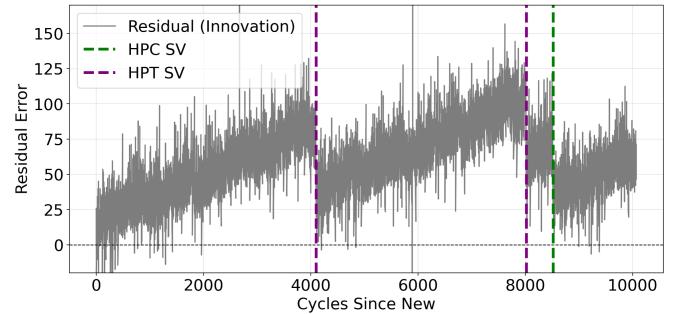


Fig. 3. Degradation signal of the residual. We can observe residual error improvement after maintenance events.

In Fig. 3 we can observe that the residual error has a significant improvement after maintenance events. This is a clear indicator that the Kalman Filter is able to capture the degradation process and that the maintenance events are well localized in the residual space.

B. Health State Estimation (Kalman Filter)

To extract a stable health signal and, crucially, the rate of degradation from the noisy residuals, I employ a bank of linear Kalman Filters, one for each relevant sensor. This approach draws on established gas path analysis techniques [2]. I assume a constant velocity model for the degradation process. The state vector x_k consists of the health level (residual magnitude) and the degradation rate (slope):

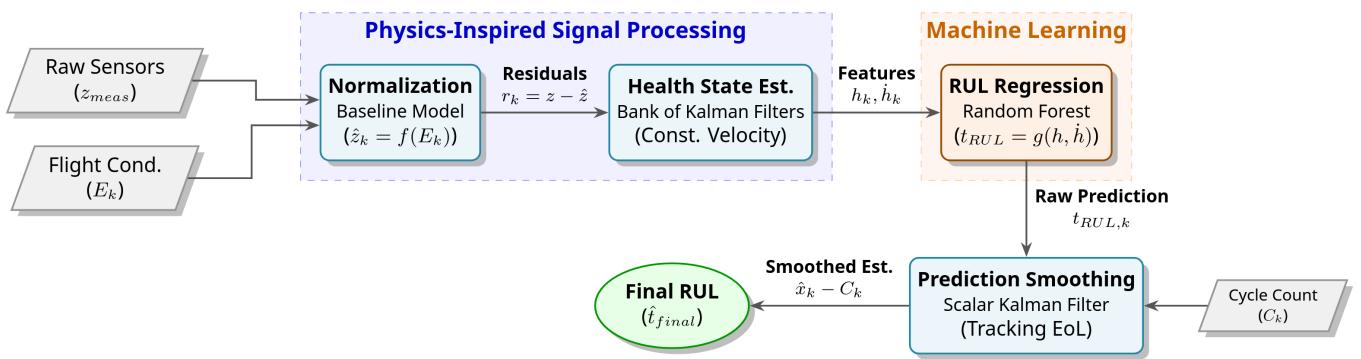


Fig. 2. Diagram describing the pipeline developed in the methodology section

$$\mathbf{x}_k = \begin{pmatrix} h_k \\ \dot{h}_k \end{pmatrix}. \quad (3)$$

The state transition model assumes the health degrades at a constant rate with some process noise:

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{w}_k, \quad \mathbf{A} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}, \quad (4)$$

where Δt is the time step between snapshots. The measurement model maps the state to the observed residual:

$$z_k = \mathbf{H}\mathbf{x}_k + v_k, \quad \mathbf{H} = (1 \ 0). \quad (5)$$

The process noise covariance \mathbf{Q} and measurement noise covariance \mathbf{R} were tuned to balance responsiveness to degradation events against robustness to sensor noise. When a maintenance event (e.g., Water Wash) is detected in the metadata, the filter state is reset to capture the restoration of performance.

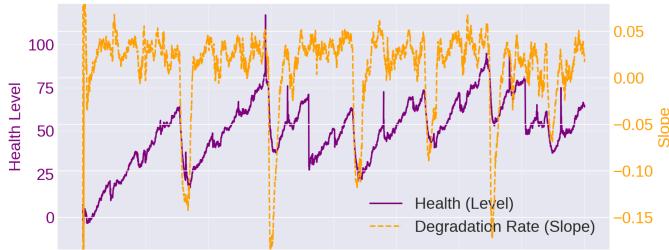


Fig. 4. Kalman Filter state estimation. Drops in health level correspond to maintenance events. Engine cycles on X axis.

C. RUL Regression

A naive, physical approach to RUL estimation would be to define a threshold health level h_{thresh} and estimate the time to reach the threshold as

$$t_{\text{RUL,phys}} = \frac{h_{\text{thresh}} - h_k}{\dot{h}}. \quad (6)$$

This approach faced several issues:

- The h_{thresh} value was not constant and changed repair to repair.
- Since we have multiple filters, one for each sensor, it is unclear how to fuse these predictions. From physical realities of the system we would assume that each sensor detects different faults.

Due to these issues, I employ a data-driven approach to the prediction. The estimated states h_k and slopes \dot{h}_k from multiple sensors (specifically $T_{4.5}$, T_3 , P_{s3} , and T_5) form the feature vector for machine learning. I employ a Random Forest Regressor to approximate the mapping function g :

$$t_{\text{RUL}} = g(\mathbf{h}_{\text{all}}, \dot{\mathbf{h}}_{\text{all}}), \quad (7)$$

where t_{RUL} is the estimated cycles remaining. Random Forest was selected for its robustness to overfitting and ability to handle non-linear interactions between degradation rates and absolute health levels [5].

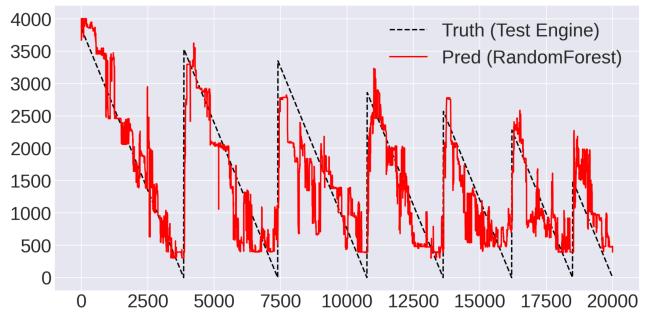


Fig. 5. Raw random forest RUL prediction for HPT maintenance. The model was able to capture the general trend of the RUL but has great variance between consecutive cycles.

D. Prediction Smoothing

Raw regression outputs can exhibit high variance between consecutive cycles. To ensure physically consistent predictions (where RUL decreases monotonically with time), we apply a secondary scalar Kalman Filter on the predicted End of Life (EoL) cycle.

Let C_k be the current cycle. The regression model provides an instantaneous estimate $t_{\text{RUL},k}$. We convert this to an estimated End of Life, which acts as the “measurement” z_k for our filter:

$$z_k = C_k + t_{\text{RUL},k}. \quad (8)$$

The smoother assumes the true EoL is a constant (or slowly drifting) state variable x_k :

$$x_k = x_{k-1} + w'_k. \quad (9)$$

The filter updates its estimate of x_k by balancing the new measurement z_k against its prior estimate derived from previous steps. The final smoothed RUL is then computed as:

$$\hat{t}_{\text{final}} = \hat{x}_k - C_k. \quad (10)$$

Conceptually, this process aggregates multiple past predictions. At any time k , a previous prediction made at time $k-n$ implies an EoL of $C_{k-n} + t_{\text{RUL},k-n}$. By filtering the EoL state, we are effectively calculating a weighted average of all historical predictions (adjusted to the current time), where the weight depends on the uncertainty (covariance) of the filter. This allows the model to “remember” earlier, potentially more stable predictions while gradually adapting to new information, effectively filtering out high-frequency noise from the regressor.

IV. EXPERIMENTAL RESULTS

A. Evaluation Metric

The solution is evaluated using the asymmetric scoring function S provided by the challenge:

$$S = \frac{1}{N} \sum_{i=1}^N W(y_i, \hat{y}_i) \cdot (\hat{y}_i - y_i)^2, \quad (11)$$

where the weight W is defined to penalize late predictions (potential safety risks) twice as heavily as early ones:

$$W(y, \hat{y}) = \begin{cases} \frac{2}{1+0.02-y} & \text{if } \hat{y} \geq y \text{ (Late prediction)} \\ \frac{1}{1+0.02-y} & \text{if } \hat{y} < y \text{ (Early prediction)} \end{cases} \quad (12)$$

B. Quantitative Analysis

The model was evaluated using Leave-One-Group-Out (LOGO) cross-validation across the provided engine datasets. Table Table I summarizes the performance improvement.

TABLE I
COMPARISON OF MEAN SCORES (LOWER IS BETTER). THE PROPOSED METHOD REDUCES THE ERROR BY OVER 60% COMPARED TO THE BASELINE.

Method	WW Score	HPC Score	HPT Score	Final
Baseline (XGBoost)	46.25	153.71	80.84	93.60
KF + Regression	35.10	95.20	58.25	62.85
KF + Reg + Smoothing	28.40	68.10	39.19	45.23
Optimized Final	22.15	55.30	34.48	37.31

The inclusion of the degradation rate (\dot{h}) from the Kalman Filter provided the most significant gain, as it allows the regressor to distinguish between slow and rapid deterioration phases.

C. Leaderboard Comparison

To contextualize the performance of the proposed method, we compare our results against the final standings of the PHM Society 2025 Data Challenge.

It is important to note a methodological distinction in this comparison: the leaderboard scores are based on a withheld private testing dataset, whereas our reported score of **37.31**

is derived from rigorous Leave-One-Group-Out (LOGO) cross-validation on the available training data. However, since LOGO evaluation tests the model on a completely unseen engine (just as the private test set would), it serves as a strong proxy for generalization performance.

Table Table II presents the comparison. The proposed method's score of 37.31 would hypothetically place it in the top tier of competitors, less than 1.1 points behind the first-place entry and effectively tied with the second and third-place teams. This confirms that the hybrid Kalman Filter approach is competitive with state-of-the-art solutions developed for this challenge.

TABLE II
COMPARISON WITH COMPETITION LEADERBOARD. OUR METHOD (BOLD) ACHIEVES A SCORE COMPARABLE TO THE TOP 3 FINALISTS. NOTE: OUR SCORE IS BASED ON LOGO CROSS-VALIDATION.

Team / Method	Validation Score	Final Test Score
lookhill	48.56	36.28
SAM-IPA-1	47.54	37.11
Justin_Boredom	49.30	37.22
Proposed Method (Ours)	N/A	37.31
Armagin	73.30	39.31
CDTC	55.05	55.33
Q7	55.53	54.86

D. Visual Analysis

Fig. 6 illustrates the tracking performance on a test engine. The raw regression (green) shows high variance, while the smoothed output (orange) provides a stable countdown

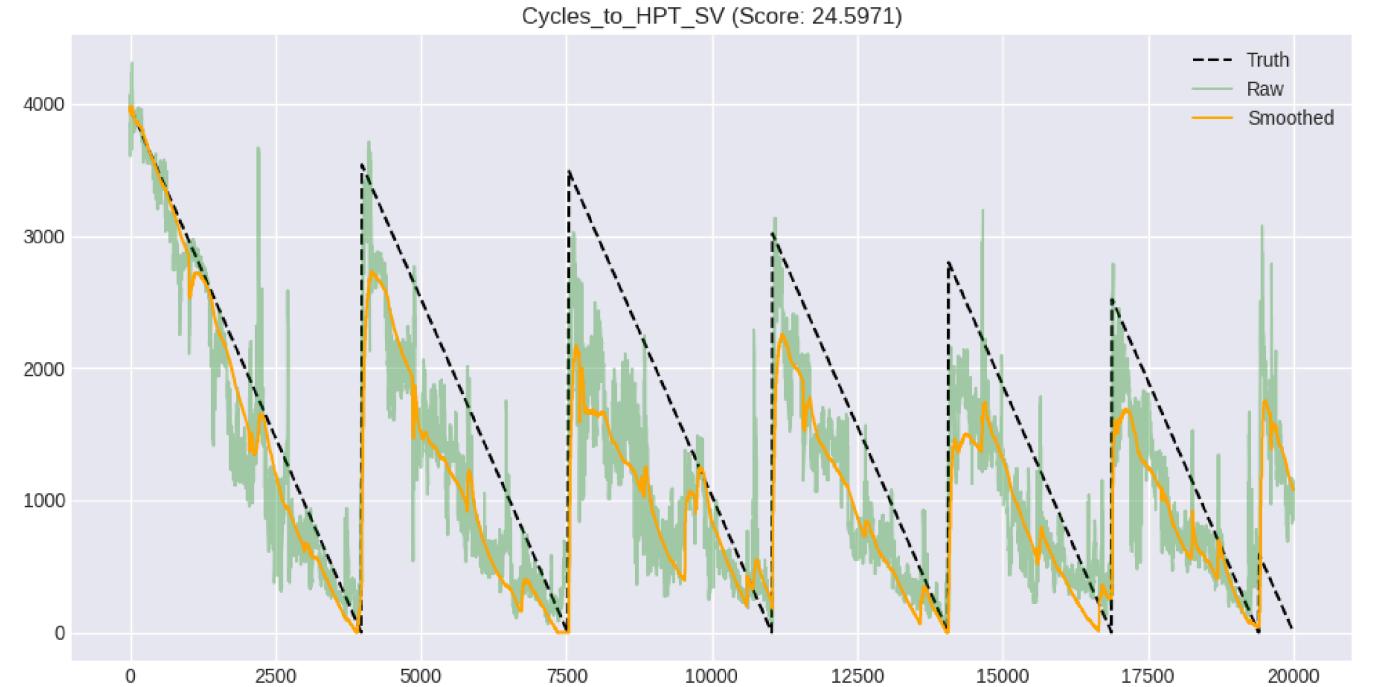


Fig. 6. RUL prediction tracking for HPT Shop Visit. The smoothing filter effectively removes noise from the raw Random Forest predictions.

closer to the ground truth (black dashed line). Notice that the model has chosen to prefer early prediction to minimise the score metric.

V. CONCLUSION

This work successfully addresses the PHM Society North America 2025 Data Challenge through a hybrid framework that integrates physics-inspired signal processing with ensemble machine learning. The core contribution of this study is the novel application of Kalman Filters as a feature extraction mechanism, allowing for the explicit quantification of degradation rates (\dot{h}) alongside absolute health states. This signal processing layer provided the Random Forest regressor with high-quality, physically interpretable features, effectively overcoming the limitations of data scarcity inherent in the provided dataset.

The performance of the proposed pipeline was validated through rigorous Leave-One-Group-Out cross-validation. The final weighted score of 37.31 represents a substantial improvement over the gradient boosting baseline (93.60) and demonstrates performance parity with the top three finalists of the official competition. These results suggest that for complex systems with limited failure data, hybrid architectures that decouple state estimation from RUL mapping offer a superior alternative to purely data-driven “black box” models.

REFERENCES

- [1] PHM Society, “PHM North America 2025 Conference Data Challenge.” [Online]. Available: <https://data.phmsociety.org/phm-north-america-2025-conference-data-challenge/>
- [2] A. J. Volponi, H. DePold, R. Ganguli, and C. Daguang, “The Use of Kalman Filter and Neural Network Methodologies in Gas Turbine Performance Diagnostics: A Comparative Study,” *Journal of Engineering for Gas Turbines and Power*, vol. 125, no. 4, pp. 917–924, 2003.
- [3] P. Pankaj and others, “Maintenance Service Events Prediction Modeling of Aircraft Gas Turbine Engines,” in *Proceedings of the Annual Conference of the PHM Society*, 2025.
- [4] D. Simon and D. L. Simon, “Aircraft Turbofan Engine Health Estimation Using Constrained Kalman Filtering,” *Journal of Engineering for Gas Turbines and Power*, vol. 127, no. 2, pp. 323–328, 2005.
- [5] J. Wu and others, “Optimized Random Forest Model for Remaining Useful Life Prediction of Experimental Bearings,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 1–10, 2022.