PHM-AP 2025 Data Challenge Technical Report: Team PrognosTech

Deepti Kunte [*1,2,+] and Vageesh Manakari[+]

[1]Siemens Industry Software NV, Interleuvenlaan 68, Leuven, Belgium
[2]KU Leuven, Department of Mechanical Engineering, Division LMSD, Heverlee, Belgium
[+]Equal contribution

1 Introduction

The PHM-AP 2025 Data Challenge is centered on predicting the progression of cutter flank wear in machining operations using high-frequency time-series data from accelerometer and acoustic emission sensors. A comprehensive description of the challenge and data is available at `https://phmap25.org/datachallenge.html#description`.

The data challenge presented several notable obstacles. Data scarcity posed a significant challenge, with only six labeled samples provided for each of the six training datasets. This limited quantity of labeled data made it difficult to model cutter flank wear progression with high accuracy and required implementation of strategies capable of generalizing from very few examples. Additionally, the sensor data from accelerometers and acoustic measurements was high-dimensional, sampled at 25.6 kHz, which increased the complexity for machine learning models and the risk of overfitting. Integrating diverse time-series data streams also posed a challenge, as the high-frequency sensor data needed to be combined with lower-frequency controller process data. This required careful data alignment and feature extraction techniques to account for differences in data granularity. Lastly, feedback from the evaluation data was restricted, as only the initial label was available for validation, making it challenging to robustly tune the model or assess its generalization performance without overfitting.

To address the challenge of data scarcity, we re-framed the regression problem by focusing on predicting relative wear differences instead of relying solely on absolute wear labels. This approach effectively expanded the set of training targets. Further, we performed comprehensive feature engineering by transforming raw sensor time series data into derived metrics such as energy and peak-to-peak values, which captured meaningful patterns in the signals while also reducing dimensionality. These combined strategies improved the robustness and effectiveness of our modeling pipeline.

2 Methodology

For each cut, the sensor and controller features (listed in Table 1) were first extracted and aggregated. For the sensor time series data, the total energy and peak-to-peak (max-min) amplitude were calculated according to Equations 1 and 2 respectively. x_i represents the sensor measurement at the i-th time step, N is the total number of samples in the cut, and T is the cut's total duration in seconds. For the controller variables, an average over all the step-wise controller measurements for each cut was used.

$$E = \frac{T}{N} \sum_{i=1}^{N} x_i^2 \tag{1}$$

$$A_{\text{p2p}} = \max(x) - \min(x) \tag{2}$$

*Corresponding author: `deepti.kunte@siemens.com`;
ORCID: `https://orcid.org/0000-0001-7857-4985`

Table 1: Features used for subsequent feature engineering

	Features
Controller	feedrate, mainSpndLoad, mainSpndSpeed, mainSpndStatus, toolSpndLoad, toolSpndSpeed, toolSpndStatus, X_load, Y_load, Z_load, A_load, B_load
Sensor	Acoustic emission and Acceleration in X, Y, and Z

The predictive modeling task was restructured to estimate total wear between any two labeled cuts i and j, where $i < j$, using the cumulative engineered features spanning those cuts. Given limited availability of wear labels, this approach effectively increased the number of training samples without introducing artificial data augmentation.

The input features consisted of a concatenation of the aggregated / engineered sensor and controller features. All features and labels were normalized to the $[0, 1]$ range using the MinMaxScaler. A random forest model was selected and its optimal hyperparameters were determined by performing cross-validation with RMSE as the evaluation metric. Each cross-validation fold trained the model on five out of the six available datasets and validated it on the remaining holdout set.

The datasets for training and inference were generated as shown in Table 2. For training, cuts $i, j \in \{1, 6, 11, 16, 21, 26\}, i < j$, while during inference cut $i = 1$, and cuts $j \in [2, 26]$.

Table 2: Structure of dataset generation for training and inference phases

Phase	**Cut i**	**Cut j**	**Feature 1**	**Feature 2**	...	**Feature N**	**Label**
Training	**i**	**j**	$\sum_{k=i+1}^{j} \text{feat1}_k$	$\sum_{k=i+1}^{j} \text{feat2}_k$	...	$\sum_{k=i+1}^{j} \text{featN}_k$	$\text{wear}_j - \text{wear}_i$
Inference	**1**	**j**	$\sum_{k=i+1}^{j} \text{feat1}_k$	$\sum_{k=i+1}^{j} \text{feat2}_k$	...	$\sum_{k=i+1}^{j} \text{featN}_k$	$\text{wear}_j - \text{wear}_1$

Since the model was trained to predict pairwise wear differences, its output at inference was $\text{wear}_j - \text{wear}_1$. To obtain the absolute predicted wear for cut j, the observed wear at cut 1 was added to the model output during inference. Finally, successive difference between the predicted wear values was taken to obtain wear estimates for each individual cut rather than cumulative wear, in line with the submission guidelines.

3 Results

The Random Forest Regressor model configured with 1,000 estimators and a maximum depth of 5 yielded the most favorable performance. On the evaluation set, the model performed reasonably well and showed stable results (Table 3) across key metrics, placing us 9th in the overall leaderboard [1].

Table 3: Summary of regression performance on test set

RMSE	**MAPE %**	**R^2**
14.603	12.709	0.797

4 Conclusions

In summary, this data challenge presented multiple interrelated obstacles, including limited labeled data, high-dimensional heterogeneous sensor streams, and strict computational constraints. By reframing the task to predict relative wear differences rather than only absolute wear values, we were able to augment our set of training targets and better leverage the available data. Furthermore, transforming raw sensor and controller time series into meaningful derived metrics enabled extraction of important patterns and reduce dimensionality, thereby enhancing the performance and generalizability of the model.

[1] We discovered a bug in our code for the final submission, which may explain the strong cross-validation scores but weaker evaluation performance. Specifically, for inference on cut j, we computed the features as $\sum_{k=1}^{j} \text{feat}_k$ instead of the intended $\sum_{k=2}^{j} \text{feat}_k$.