A Comparative Analysis of Machine Learning Models for Aircraft Engine Remaining Useful Life Prediction

1st Muhammad Fitran Ramadhan
Computer Engineering Study Program,
School of Electrical Engineering
Telkom University
Bandung, Indonesia
fitranhp@gmail.com

2nd Fahrial Al Akbar Warsito Putra
Computer Engineering Study Program,
School of Electrical Engineering
Telkom University
Bandung, Indonesia
fahrialalakbarwp@gmail.com

3rd Nabil De Nozyra
Computer Engineering Study Program,
School of Electrical Engineering
Telkom University
Bandung, Indonesia
nabil.denozyra@gmail.com

4th Surya Michrandi Nasution
Computer Engineering Study Program,
School of Electrical Engineering
Telkom University
Bandung, Indonesia
michrandi@telkomuniversity.ac.id

5th Reza Rendian Septiawan
Computer Engineering Study Program,
School of Electrical Engineering
Telkom University
Bandung, Indonesia
zaseptiawan@telkomuniversity.ac.id

***Abstract*—Accurate prediction of Remaining Useful Life (RUL) for aircraft turbofan engines enables predictive maintenance, enhancing operational safety, and reducing costs. This paper presents a comparative analysis of machine learning models for RUL prediction using the benchmark NASA C-MAPSS dataset. We systematically evaluate the performance of three distinct regressors—Linear Regression, XGBoost, and K-Nearest Neighbors—across twelve unique preprocessing pipelines. These pipelines combine four denoising methods (Moving Average, EWMA, Wavelet, and no denoising) and three feature scaling techniques (Standardization, Min-Max, and Robust Scaling). Our results demonstrate that the data preprocessing strategy is a more critical determinant of success than the choice of model. The Wavelet denoising technique, intended to improve the signal, was found to be universally detrimental, severely degrading performance and increasing the average Root Mean Squared Error (RMSE) to approximately 80. Conversely, the highest accuracy was achieved using either a simple Moving Average (MA) filter (average RMSE ≈ 56.5) or no denoising at all (average RMSE ≈ 57.5). Regarding model selection, non-linear models proved superior for complex scenarios. On the multi-fault dataset (FD003), XGBoost was the most robust, achieving an RMSE of 46.6, significantly outperforming Linear Regression (RMSE ≈ 68.2) and KNN (RMSE ≈ 67.3). This study concludes that modern machine learning models are highly robust to raw sensor noise and that validating preprocessing steps is essential to prevent the inadvertent removal of critical prognostic information.**

Keywords*—*prognostics and health management (PHM), RUL, machine learning, data preprocessing, predictive maintenance

I. Introduction

The reliability and safety of aircraft engines are paramount in the aviation industry. Historically, maintenance strategies were often reactive or based on fixed schedules, leading to either premature replacement of healthy components or, conversely, unexpected failures [1]. The modern paradigm has shifted towards PHM, a framework that enables predictive maintenance [2]. By continuously monitoring a system's health and predicting its future state, PHM promises considerable cost savings by avoiding unscheduled maintenance, increasing equipment usage, and significantly improving operational safety [3]. This proactive approach allows maintenance to be scheduled precisely when needed, maximizing the operational life of components while minimizing the risk of in-flight incidents.

At the core of any advanced PHM system lies the task of prognostics, which is exclusively defined as the estimation of a component's RUL [4]. RUL is typically estimated in units of time, such as flight hours or operational cycles [5]. An accurate RUL prediction provides decision-makers with critical information, allowing them to adjust operational loads to prolong component life or to initiate the logistics process for a smooth transition from faulty to functional equipment. The primary challenge is to develop robust models that can accurately forecast RUL using historical and real-time sensor data from the engine's operation.

To effectively train and validate data-driven prognostic models, a dataset with complete run-to-failure trajectories is essential. National Aeronautics and Space Administration (NASA) developed the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset [6]. This dataset is ideal for our research as it provides multiple multivariate time series generated from a high-fidelity thermodynamic model of a turbofan engine. Each time series represents the operational history of a single engine from a fleet, starting with varying degrees of normal initial wear and ending in system failure. The data for each operational cycle includes three operational settings and twenty-one sensor measurements, simulating the information typically available for monitoring. The training subset contains the full degradation path to failure, providing the ground truth RUL necessary for supervised machine learning. The objective is to use this data to predict the RUL of the test set engines, whose time series end prior to failure. The availability, realism, and completeness of the C-MAPSS data make it the definitive choice for this comparative study.

The objective of this paper is to conduct a comparative analysis of machine learning models for predicting the RUL of aircraft turbofan engines using the benchmark C-MAPSS dataset. Systematic evaluation the performance of three distinct algorithms: Linear Regression as a simple linear model and performance baseline, XGBoost as a powerful ensemble model, and K-Nearest Neighbors as a non-parametric model. Furthermore investigating the impact of various denoising and scaling methods. The primary contribution of this work is to provide a clear comparison of these model-preprocessing combinations, offering insights into their effectiveness in handling the varying complexities presented by the dataset.

II. Literature Review

The literature review chapter provides a survey of existing research and methodologies relevant to the prediction of RUL in aircraft turbofan engines. By examining foundational studies and recent innovations, this review establishes the theoretical framework for this study's comparative analysis of machine learning models and preprocessing strategies. Understanding the progression and gaps in prior work for contextualizing the methodology and highlighting the contributions of this research.

A. Dataset: NASA C-MAPSS

This research uses C-MAPSS dataset. The data was generated using a high-fidelity model of a commercial turbofan engine of the 400 kN thrust class (originally specified as 90,000 lbf). The dataset consists of multiple multivariate time series, where each time series represents the full operational life of a single engine from a larger fleet. Each engine begins with different, unknown degrees of initial wear and manufacturing variations, which are considered normal operational conditions. In the training sets, a fault develops and progresses until system failure, providing the complete ground truth RUL for model training. In the test sets, the time series ends a number of cycles prior to failure, and the objective is to predict this remaining number of cycles. The data is divided into four distinct subsets, each with increasing complexity. The simplest subset, FD001, involves a single operating condition and one fault mode of High-Pressure Compressor (HPC) degradation. The complexity increases in FD002, which features six operating conditions with the same single fault mode. FD003 returns to a single operating condition but introduces a second fault mode (Fan degradation). FD004 represents the most complex scenario with six operating conditions and two fault modes. Each row in the data files represents a single operational cycle and contains 26 columns: the unit number, time in cycles, three operational settings, and twenty-one sensor measurements.

B. Data Preprocessing

This stage involves two processes: noise reduction (denoising) and feature scaling. The overall goal is to provide the models with a cleaner, standardized signal, which helps them learn the true relationship between sensor readings and RUL, leading to more accurate and generalizable predictions.

1) Denoising Methods

The first preprocessing stage involves noise reduction, or denoising. The C-MAPSS dataset is contaminated with sensor noise, and this stage is expected to improve model performance by aiming to remove irrelevant noise and isolate the underlying degradation trend.

The first approach, No Denoising, serves as a baseline. This method involves feeding the raw, unfiltered sensor data directly into the scaling process. By evaluating this pipeline, it is possible to determine whether denoising provides any tangible benefit or if the models are inherently robust enough to handle the sensor noise present in the C-MAPSS data.

The Moving Average (MA) filter is a common time-series smoothing technique applied to reduce short-term fluctuations [7]. It operates by calculating the mean of the sensor values over a fixed-size sliding window. This process effectively smooths out random noise and highlights the longer-term degradation trend. The primary assumption is that the underlying degradation signal is a low-frequency component, while the noise is high-frequency.

The Exponentially Weighted Moving Average (EWMA) filter refines the simple MA approach by assigning greater importance to more recent data points [7]. It calculates the average by applying an exponential decay factor, meaning the influence of older observations decreases over time. This method is particularly useful if the most recent sensor readings are more predictive of the current health state than older readings within a window.

Finally, Wavelet Denoising was employed as a more sophisticated signal processing technique. This method decomposes the signal into different frequency scales, allowing for the separation of the underlying signal from noise [8]. It operates by performing a wavelet transform, "thresholding" the detail coefficients (which are assumed to be noise), and then reconstructing the signal. This is expected to preserve sharp features of the degradation signal while removing noise across multiple frequencies.

2) Feature Preprocessing

The second stage is feature scaling, which standardizes the range of the sensor data to ensure all features contribute equally during model training.

Standardization, or Z-score normalization, was the first method tested. It transforms the data by subtracting the mean and dividing by the standard deviation (calculated from the training set) [9]. This process centers the features around a mean of zero with a unit variance. Standardization is effective for models that assume a Gaussian distribution of the input features, such as Linear Regression.

Min-Max Scaling normalizes features by scaling them to a fixed range, typically from 0 to 1 [9]. The transformation is performed by subtracting the minimum value and dividing by the range (maximum minus minimum) of the feature, as observed in the training data. This method is highly effective for algorithms like K-Nearest Neighbors, which are distance-based and sensitive to the magnitude of features.

Robust Scaling was chosen for its resilience to outliers. Instead of the mean and standard deviation, this method uses the median and the interquartile range (IQR) for scaling [10]. By subtracting the median and dividing by the IQR, it produces a transformation that is not significantly influenced by a few very large or very small sensor readings, which are common in noisy operational data.

By combining the four denoising options (including no denoise) with the three scaling methods, this study systematically evaluates a total of 12 distinct preprocessing pipelines for each machine learning model to achieve the most effective combination for RUL prediction.

C. Regressor Models

There will be three machine learning models with varying levels of complexity to evaluate their performance on the RUL prediction task.

Linear Regression was chosen to represent simple, parametric statistical models and to serve as a performance baseline [11]. This model works by finding an optimal linear relationship between the input features and the target RUL. Its simplicity makes it highly interpretable and computationally efficient. By evaluating its performance first, it is possible to gauge the inherent difficulty of the prediction task and determine the added value of more complex approaches.

XGBoost (Extreme Gradient Boosting) was selected to represent the class of powerful, non-linear, tree-based ensemble models [12]. XGBoost builds a highly accurate predictor by sequentially training a series of decision trees, where each new tree is designed to correct the errors made by the previous ones. Such models are known for their state-of-the-art performance on structured, tabular data, as they can automatically capture complex interactions and non-linear relationships between the sensor features, which are expected in engine degradation data.

K-Nearest Neighbors (KNN) was chosen to represent instance-based, non-parametric regressors [13]. Unlike the other models, KNN makes no assumptions about the underlying data distribution. It is a "lazy learning" algorithm that predicts the RUL of a given data point by finding the 'k' most similar instances in the training set and averaging their RUL values. This approach provides a completely different perspective, as its performance relies on the local structure of the data, contrasting with the global linear assumptions of Linear Regression and the hierarchical structure of XGBoost.

D. Evaluation Metrics

To quantitatively assess and compare the performance of the prognostic models, several standard regression metrics will be used. The Root Mean Squared Error (RMSE) was used to measure the average magnitude of the errors while giving a higher weight to large errors. The Mean Absolute Error (MAE) provides a more direct interpretation of the average error magnitude. Finally, the R-squared (R^2) metric, or coefficient of determination, was used to evaluate the proportion of variance in the RUL that the model could successfully explain [14].

III. Methodology

This chapter presents a systematic approach for predicting the RUL of aircraft turbofan engines, a key task in enabling predictive maintenance and enhancing aviation safety. This methodology uses the NASA C-MAPSS dataset, which contains extensive run-to-failure sensor data from turbofan engines under varying operational conditions, making it ideal for supervised learning models. The preparation of this data includes preprocessing steps such as denoising and feature scaling to mitigate sensor noise and normalize inputs, thereby enhancing model learning. The study evaluates three distinct machine learning models: Linear Regression, XGBoost, and K-Nearest Neighbors, chosen for their diversity in complexity and modeling capability, assessing their performance across multiple preprocessing pipelines.

Identifying the most reliable combination of models and data preprocessing strategies that can accommodate the non-linear, high-dimensional sensor data often contaminated with noise and subject to diverse operating conditions. This approach aligns with recent advances in the field, such as domain adaptation and deep learning models, which have shown promising results in handling the challenges associated with engine degradation data [15, 16]. The comprehensive evaluation considers the robustness, accuracy, and interpretability of the models, providing guidance for developing more effective prognostics systems capable of operating reliably in real-world environments. This thorough experimental framework aims to identify optimal combinations of models and preprocessing techniques to accurately estimate RUL, addressing the complexities of engine degradation signals and operational variability.

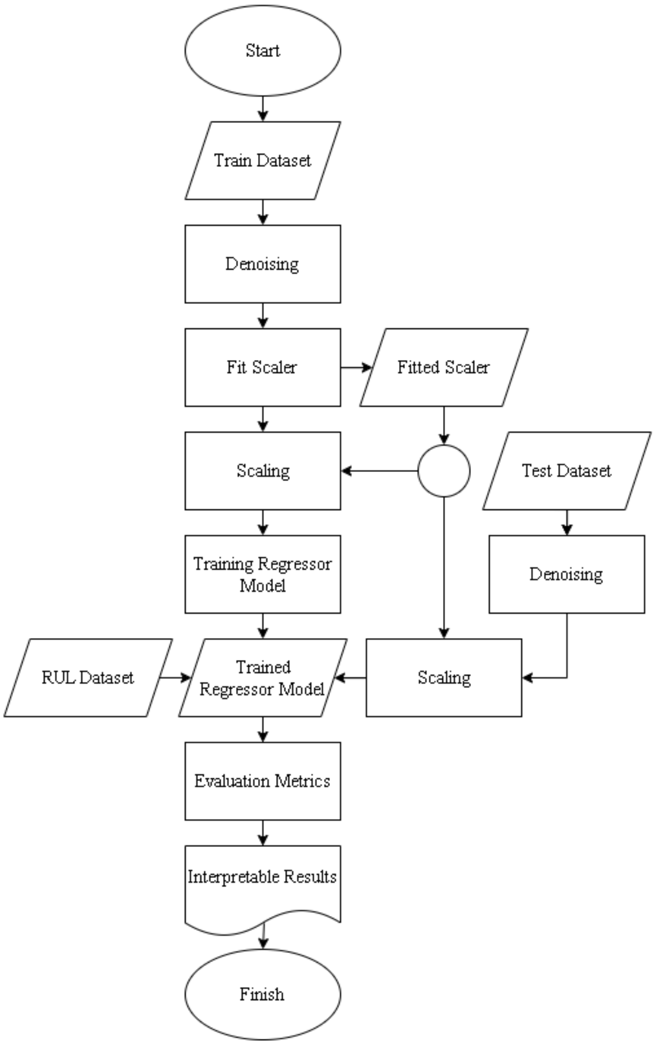

Fig. 1. Experimental workflow for model training and evaluation.

Fig. 1 outlines the structured experimental pipeline used for training, testing, and evaluating the machine learning models. The workflow begins by applying a denoising filter to the Train Dataset. Following this, a feature scaler is fitted exclusively to this denoised training data to learn its statistical properties, such as its mean and variance or min-max range. This "Fitted Scaler" is then used to transform the training data, which, along with the ground truth RUL Dataset, is used to train the regressor model.

For the testing phase, the Test Dataset undergoes the identical preprocessing sequence to ensure consistency and prevent data leakage [17]. It is first passed through the same denoising filter. Then, crucially, it is scaled using the 'Fitted Scaler' that was learned only from the training data. The Trained Regressor Model uses this preprocessed test data to make predictions. Finally, these predictions are compared against the ground truth RUL values using Evaluation Metrics to quantify model performance and generate the final interpretable results.

IV. Results

This section presents the performance evaluation of the three machine learning models (Linear Regression, XGBoost, and K-Nearest Neighbors) across the twelve distinct preprocessing pipelines. The analysis is segmented into the aggregate impact of preprocessing techniques and a detailed comparison of model performance against the four C-MAPSS datasets.

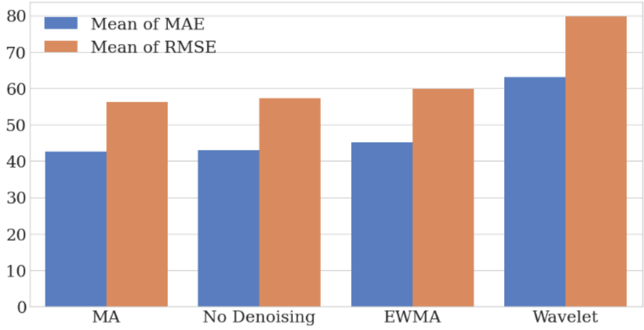

Fig. 2. Mean of MAE and RMSE grouped by denoising method.

The experimental results first reveal that the choice of the preprocessing pipeline, particularly the denoising method, is a substantial determinant of model performance. Fig. 2 illustrates the aggregate impact of each denoising method. The wavelet denoising technique, a method designed to separate signals from noise across multiple frequencies, consistently and significantly degraded model performance. This counter-intuitive result suggests incompatibility with the C-MAPSS data. The wavelet filter might misinterpret crucial, high-frequency degradation signatures such as subtle component rubbing or vibrations that signal the start of a fault which are essential for prognostics. The resulting "clean" signal, while smoother, is information-poor, leading to a substantial increase in prediction error (RMSE $\approx$ 80).

Conversely, the strong performance of the 'No Denoising' baseline (RMSE $\approx$ 57.5), which was nearly identical to the simple MA filter (RMSE $\approx$ 56.5), is highly instructive. This suggests that the sensor noise present in the C-MAPSS data is not the primary challenge for the models. Non-linear models like XGBoost, in particular, appear capable of isolating the underlying degradation trend despite the noise. The slight benefit of the MA filter implies that smoothing only the most high-frequency, short-term fluctuations is beneficial, whereas the more aggressive, multi-scale filtering of the wavelet is detrimental.

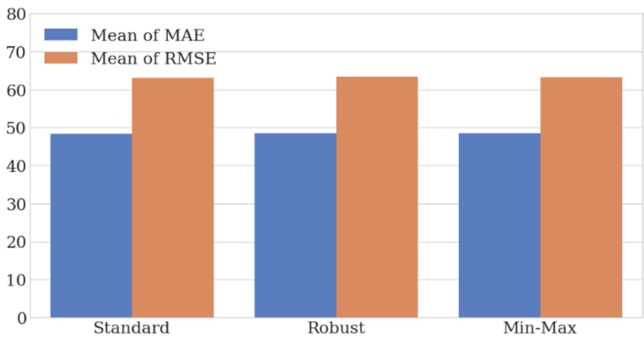

Fig. 3. Mean of MAE and RMSE grouped by feature scaling method.

In contrast, Fig. 3 shows that the choice of feature scaling method (Standardization, Robust, or Min-Max) had a negligible impact on aggregate performance across all models. This is logical, as tree-based models like XGBoost are largely insensitive to feature magnitude. For Linear Regression and KNN, the primary goal is simply to normalize the features to a comparable scale; the specific method used appears to be of low importance for this dataset.

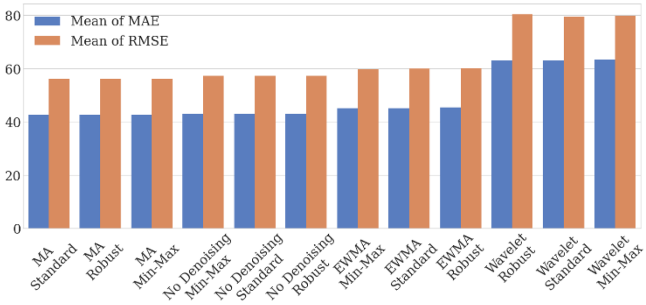

Fig. 4. Mean of MAE and RMSE for all 12 preprocessing pipelines.

Fig. 4 visually confirms these findings, showing a clear separation in performance. The three worst-performing pipelines are all those that incorporate Wavelet denoising, supporting the conclusion that the Wavelet filtering itself, not its interaction with scaling, is the root cause of the performance degradation.

While preprocessing was a dominant factor, the choice of model became increasingly important as the dataset complexity—defined by operational conditions and fault modes—increased. Table I. Best performance metrics and pipelines by model for each dataset. Fig. 4. Mean of RMSE achieved by each model across the four datasets. Fig. 5. Mean of MAE achieved by each model across the four datasets. Fig. 6. Best of R^2 score achieved by each model across the four datasets. Figs. 5, 6, and 7 , which plot the best result for each model on each dataset (with full data in Table I), reveal key insights into each model's capabilities.

Detailing the optimal performance metrics (RMSE, MAE, and R^2) achieved by each of the three models—Linear Regression, XGBoost, and KNN—on each of the four C-MAPSS datasets, as shown in Table I. More than just listing the scores, it specifies the exact preprocessing pipeline (the specific Denoising and Scaling method) that allowed each model to achieve its best result for that particular dataset. Generally, the table illustrates how model performance (e.g., XGBoost's lower error on FD003) and the most effective preprocessing strategy (e.g., the frequent appearance of "No Denoising" or "MA") shift in response to the increasing complexity of the datasets, from the simple FD001 to the multi-fault, multi-condition FD004.

Fig. 5 illustrates the models' comparative robustness to increasing data complexity, with RMSE penalizing large errors more heavily. On the simple FD001 dataset, all models perform comparably. However, on FD003 (two faults, one condition), a clear divergence occurs: the RMSE for Linear Regression and KNN increases significantly (to $\approx$ 68), while XGBoost's RMSE remains low ($\approx$ 46.6).

TABLE I. BEST PERFORMANCE METRICS AND PIPELINES BY MODEL FOR EACH DATASET

Dataset	Model	Denoising	Scaling	RMSE	MAE	R^2
FD001	Linear Regression	MA	Min-Max	42.68624436359004	32.93145879282506	0.4761609868045232
	XGBoost	No Denoising	Min-Max	43.68458175659180	33.24300384521484	0.4513714909553528
	KNN	No Denoising	Min-Max	43.12798285843578	32.84511122710628	0.4652629946702862
FD002	Linear Regression	MA	Standard	46.06667030908218	36.59298849778088	0.4798441138732038
	XGBoost	MA	Robust	43.92177581787109	34.34372711181640	0.5271540880203247
	KNN	No Denoising	Min-Max	45.05351551355199	34.94426867004902	0.5024723276584605
FD003	Linear Regression	No Denoising	Robust	68.22052775842812	50.94343562984176	0.3439246589723204
	XGBoost	MA	Robust	46.65368652343750	35.69073104858398	0.3742600083351135
	KNN	No Denoising	Robust	67.34917015299473	48.62024803348013	0.3605772808062044
FD004	Linear Regression	MA	Standard	72.28630536530142	54.39470366259392	0.3848888081278651
	XGBoost	No Denoising	Robust	71.10757884290423	53.11040599619724	0.4047857008127592
	KNN	No Denoising	Robust	71.10757884290423	53.11040599619724	0.4047857008127592

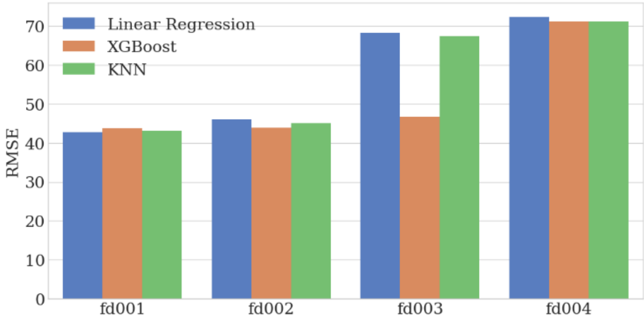

Fig. 5. Mean of RMSE achieved by each model across the four datasets.

This demonstrates that Linear Regression cannot model the non-linear interaction of two faults, and KNN's instance-based logic fails when the feature space becomes complex. On FD004, all models converge to a high RMSE ($\approx$ 71-72), indicating the compounded challenge of multiple faults and multiple operating conditions exceeds the predictive capacity of all tested models.

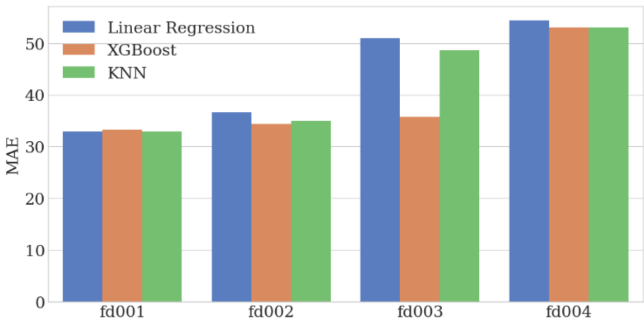

Fig. 6. Mean of MAE achieved by each model across the four datasets.

As shown in Fig. 6, the MAE closely mirrors the trends observed in the RMSE analysis from Fig. 5. This consistency supports the previous findings, particularly that the simpler models (Linear Regression, KNN) struggle significantly more than XGBoost when faced with the multi-fault dynamics of dataset FD003.

Fig. 7 shows the R^2 score, or the proportion of variance in the RUL that each model could explain. While XGBoost consistently achieved the highest R^2 scores, particularly on FD002 ($R^2 \approx 0.53$) and FD003 ($R^2 \approx 0.37$) , the most notable finding is the sharp drop in performance for all models on the multi-fault datasets (FD003 and FD004).

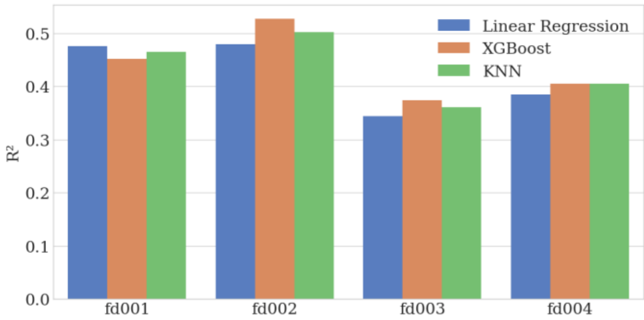

Fig. 7. Best of R^2 score achieved by each model across the four datasets.

This indicates that even the best-performing model failed to explain a large portion of the RUL variance in these complex scenarios, highlighting the inherent difficulty of this prognostic task.

V. CONCLUSSION

This systematic comparison of machine learning models and preprocessing pipelines for RUL prediction found that the data preprocessing strategy, particularly denoising, had a more substantial impact on performance than the choice of algorithm. Counter-intuitively, the theoretically advanced wavelet denoising technique was consistently detrimental to predictive accuracy. We conclude that this filter, in its application here, removes critical high-frequency prognostic signatures that it misinterprets as noise.

This underscores an important lesson for PHM: aggressive data filtering can be counter-productive, and validation is essential to ensure that prognostic information is not "cleaned" away. The strong performance of models with minimal (MA) or no denoising suggests that modern algorithms, particularly XGBoost, are robust to the raw sensor noise present in the data. While all models performed comparably on simple, single-fault scenarios , XGBoost emerged as the most effective model as data complexity increased, demonstrating a superior ability to capture the non-linear interactions of multiple fault modes.

In conclusion, this research indicates that the optimal strategy is not simply to pair the most complex filter with the most complex model. Instead, the most effective approach is to use a robust, non-linear model like XGBoost and pair it with a minimalist preprocessing pipeline that trusts the model to isolate the degradation trend from the inherent sensor noise.

Future research should build on these findings in several key directions. First, more advanced model architectures should be explored; while XGBoost was effective, Recurrent Neural Networks (RNNs) like LSTM or GRUs may better capture the temporal dependencies inherent in degradation data. Second, rather than dismissing wavelet denoising entirely, a deeper investigation is warranted by tuning its parameters, such as wavelet selection, decomposition levels, and thresholding functions, to potentially isolate and retain the high-frequency prognostic information that was lost in this study. Finally, this analysis could be expanded by comparing these results against other advanced denoising techniques, like Kalman filters, and alternative feature scaling methods to create an even more robust and comprehensive preprocessing strategy.

REFERENCES

[1] T. van der Weide, Q. Deng, and B. F. Santos, "Robust long-term aircraft heavy maintenance check scheduling optimization under uncertainty," *Computers & Operations Research*, vol. 141, art. 105667, 2022, doi: 10.1016/j.cor.2021.105667.

[2] C. Huang, S. Bu, H. H. Lee, C. H. Chan, S. W. Kong, and W. K. C. Yung, "Prognostics and health management for predictive maintenance: A review," *Journal of Manufacturing Systems*, vol. 75, pp. 78–101, 2024. doi: 10.1016/j.jmsy.2024.05.021.

[3] L. Polverino *et al.*, "Machine learning for prognostics and health management of industrial mechanical systems and equipment: A systematic literature review," *International Journal of Engineering Business Management*, vol. 15, 2023. doi: 10.1177/18479790231186848.

[4] S. Ochella, M. Shafiee, and F. Dinmohammadi, "Artificial intelligence in prognostics and health management of engineering systems," *Engineering Applications of Artificial Intelligence*, vol. 108, art. 104552, 2022, doi: 10.1016/j.engappai.2021.104552.

[5] A. Peringal, M. B. Mohiuddin, A. G. Haddad, and P. K. Muthusamy, "Reliable Prediction of Remaining Useful Life for Aircraft Engines: An LSTM-Based Approach With Conservative Loss Function," in *AIAA 2025-1909*, 2025, doi: 10.2514/6.2025-1909.

[6] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation," in Proc. 1st International Conference on Prognostics and Health Management (PHM08), Denver, CO, 2008.

[7] V. Lotysh, L. Gumeniuk, and P. Humeniuk, "Comparison of the effectiveness of time series analysis methods: SMA, WMA, EMA, EWMA, and Kalman filter for data analysis," *IAPGOŚ*, no. 3, pp. 71–74, Sep. 2023, doi: 10.35784/iapgos.3652.

[8] X. Li and L. Zeng, "Research on signal denoising method based on improved wavelet threshold," in *EITCE '23: Proceedings of the 2023 International Conference on Electronic Information Technology and Computer Engineering*, Xiamen, China, 2024, pp. 857–861, doi: 10.1145/3650400.3650545.

[9] Y.-S. Kim *et al*., "Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models," *Sustainable Cities and Society*, vol. 118, art. 105570, 2025, doi: 10.1016/j.scs.2024.105570.

[10] A. Yang *et al*., "tanh As a robust feature scaling method in training deep learning models with imbalanced data," *Pattern Recognition*, vol. 167, art. 111746, 2025, doi: 10.1016/j.patcog.2025.111746.

[11] E. M. Flores Hinostroza, D. J. Mendoza, M. Navarro Cejas, and E. P. Palacios Trujillo, "Linear regression model to predict the use of artificial intelligence in experimental science students," *International Electronic Journal of Mathematics Education*, vol. 20, no. 1, art. em0807, 2025, doi: 10.29333/iejme/15736.

[12] M. Alizamir *et al*., "An interpretable XGBoost-SHAP machine learning model for reliable prediction of mechanical properties in waste foundry sand-based eco-friendly concrete," *Results in Engineering*, vol. 25, art. 104307, 2025, doi: 10.1016/j.rineng.2025.104307.

[13] M. Seoane Santos, P. H. Abreu, S. Wilk, and J. Santos, "How distance metrics influence missing data imputation with k-nearest neighbours," *Pattern Recognition Letters*, vol. 136, pp. 111–119, 2020, doi: 10.1016/j.patrec.2020.05.032.

[14] H. Khoshvaght, R. R. Permala, A. Razmjou, and M. Khiadani, "A critical review on selecting performance evaluation metrics for supervised machine learning models in wastewater quality prediction," *Journal of Environmental Chemical Engineering*, vol. 13, no. 6, art. 119675, 2025, doi: 10.1016/j.jece.2025.119675.

[15] Y. Wang, M. Ragab, Y. Hou, Z. Chen, M. Wu, and X. Li, "Deep domain adaptation for turbofan engine remaining useful life prediction: Methodologies, evaluation and future trends," *arXiv e-prints*, Oct. 2025. doi: 10.48550/arXiv.2510.03604.

[16] O. Asif *et al.*, "A deep learning model for remaining useful life prediction of aircraft turbofan engine on C-MAPSS dataset," *IEEE Access*, vol. 10, pp. 95425-95440, 2022, doi: 10.1109/ACCESS.2022.3204028.

[17] S. R. Kumar and V. K. Vatsavayi, "Performance analysis of machine learning techniques for server health monitoring using time series data against DDOS attacks," IEEE Access, vol. 13, pp. 53321-53346, Apr. 2025, doi: 10.1109/ACCESS.2025.3553558.