Remaining Useful Life (RUL) Prediction of Turbofan Engines Using Explainable AI: Random Forest, LIME, and SHAP

Abhay Raj Yadav
E22CSEU1276@bennett.edu.in
Bennett University

Vanipally Rahul Reddy
E22CSEU1274@bennett.edu.in
Bennett University

Ayush Sharma
E22CSEU1271@bennett.edu.in
Bennett University

Adnan Hasan
E22CSEU1269@bennett.edu.in
Bennett University

***Abstract*—Predictive maintenance has become a critical component of modern aerospace engineering, ensuring operational safety, improving efficiency and minimizing costs. This study focuses on predicting the Remaining Useful Life (RUL) of turbofan engines using the NASA C-MAPSS data set. Leveraging the robustness of a Random Forest model, we address the inherent complexity of engine degradation under various operating conditions. The framework also incorporates advanced explainability techniques, namely Local Interpretable ModelAgnostic Explanations (LIME) and Shapley Additive explanations (SHAP), to clarify the contributions of specific features to model predictions.**

The research evaluates the proposed methodology on four subsets of the C-MAPSS data set (FD001–FD004), each representing different levels of operational complexity and failure modes. The results demonstrate the effectiveness of the approach, reaching R^2 values of up to 0.6560 on simpler data sets. Challenges are observed when handling data sets with more complex failure modes, such as the cases of FD003 and FD004. Additionally, interpretability analyzes provide valuable information on the behavior of critical sensors and operational configurations that influence RUL predictions. This study highlights the importance of combining predictive accuracy with interpretability, providing actionable insights for decision making in aerospace maintenance.

In a novel approach, the PySpark distributed computing framework was leveraged to handle the NASA C-MAPSS dataset with unmatched scalability and efficiency. Tasks such as feature scaling, normalization, splitting datasets, and handling missing values were simplified using PySpark's DataFrame API, allowing for smooth manipulation of high-dimensional data and a significant reduction in computation times thanks to to parallel execution. The Random Forest model was trained on all subsets (FD001 to FD004), while Decision Tree and Gradient Boosting focused on FD002, taking advantage of its moderate complexity and optimal performance for detailed evaluation. PySpark's MLlib library facilitated optimized model evaluation, hyperparameter tuning, and pipeline management for large-scale machine learning tasks.

Advanced interpretability techniques, such as Local Independent Model Explanations (LIME) and Shapley Additive Explanations (SHAP), clarified the contributions of features. The results demonstrated the effectiveness of this hybrid framework, which combines the scalability of PySpark with interpretable insights, achieving R^2 values of 0.6560 for FD002. These findings offer actionable insights into engine health and critical sensor behaviors, enabling better predictive maintenance strategies.

Keywords— Predictive maintenance, Remaining Useful Life (RUL), Turbofan engines, NASA C-MAPSS dataset, Apache spark, Random Forest, Engine degradation, LIME (Local Interpretable Model-Agnostic Explanations), SHAP (Shapley Additive Explanations), Model interpretability, and R^2 values. These keywords highlight the core topics and methodologies discussed in your study, focusing on predictive maintenance techniques, the data used, and the explainability of the mode

I. INTRODUCTION

The aerospace industry is at the forefront of adopting predictive maintenance strategies to address the challenges posed by traditional maintenance paradigms. Conventional maintenance approaches, such as time-based or reactive strategies, are often inefficient and lead to excessive operating costs or unexpected failures[3]. In contrast, predictive maintenance leverages advanced data analytics and machine learning to anticipate equipment failures, enabling timely interventions and reducing downtime.

The Remaining Useful Life (RUL) of an engine component is a critical metric in predictive maintenance. Accurate RUL predictions allow maintenance teams to optimize schedules, reduce unnecessary repairs, and improve safety. However, the complex interplay between operating conditions, environmental factors, and component degradation presents significant challenges to RUL prediction models [6]. The NASA C-MAPSS data set has emerged as a benchmark for developing and testing such models, offering simulated data of turbofan engines under various conditions and failure modes.

Machine learning, particularly deep learning, has gained popularity for RUL prediction due to its ability to model complex, non-linear relationships. However, deep learning models often lack interpretability, limiting their practical applicability in high-risk industries such as aerospace[7]. This study addresses this limitation by using a Random Forest model, a tree-based ensemble learning method known for its transparency and reliability. To further improve the interpretability of the model, we integrated LIME and SHAP,[6,9] which provide insights into the model's decision-making process and identify key factors that influence predictions. This dual approach ensures that the model is not only accurate, but also reliable and applicable.

II. RELATED WORK

The field of RUL prediction has evolved significantly over the past decade, driven by advances in machine learning and the availability of large-scale data sets. Initial approaches relied on statistical models and domain-specific physical simulations to estimate component degradation. Although these methods provided valuable information, they often struggled with scalability and the ability to generalize across different operating conditions[10].

With the advent of data-driven methods, machine learning models such as Support Vector Machines (SVM), Random Forest and Gradient Boosting Machines have become popular for RUL prediction. These models excel at handling highdimensional data and deliver interpretable results through feature importance scores. More recently, deep learning architectures such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN) have achieved state-of-the-art performance by capturing temporal dependencies and complex interactions between features. However, its "black box" nature has raised concerns about transparency and reliability, particularly in safety-critical applications[10].

To address these challenges, researchers have turned to explainability techniques that provide insights into model predictions. LIME and SHAP are two prominent methods that have been widely adopted in various fields. LIME explains individual predictions by approximating the model locally with interpretable surrogates, while SHAP provides a global understanding of feature contributions using game theory principles[6]. By integrating these techniques, we aim to improve the interpretability of our Random Forest model, ensuring that predictions are both accurate and actionable.

III. DATASET AND PREPROCESSING

A. Dataset Description

The NASA C-MAPSS data set is a widely used benchmark for RUL prediction in aerospace engineering. Simulates the performance of turbofan engines under various operating conditions and failure scenarios. The data set consists of four subsets (FD001 to FD004), each characterized by different combinations of operating conditions and failure modes:

FD001: A single operating condition with one failure mode.

FD002: Various operating conditions with a failure mode.

FD003: A single operating condition with two failure modes.

FD004: Various operating conditions with two failure modes.

Each subset includes:

Operational settings: Three continuous variables representing environmental and operational conditions.

Sensor Measurements: Twenty-one continuous variables that capture engine health and degradation patterns.

Motor unit numbers and cycles: Each motor unit is monitored over time, with its cycles serving as a temporal dimension for RUL prediction.

The subsets differ in complexity, with FD001 representing a simpler scenario and FD004 capturing more realistic and challenging conditions.

B. Preprocessing Steps

Preprocessing plays a critical role in preparing data for model training and evaluation. The following steps were applied to the C-MAPSS data set:

RUL Calculation: For each motor, the Remaining Useful Life is calculated as the difference between the maximum cycle and the current cycle. This ground truth label serves as the target variable for training the model.

Data Cleansing: Redundant sensors are removed, such as Sensors 22 and 23, which contain constant values across all instances. This reduces noise and improves computational efficiency.

Normalization: All sensor and operational characteristics are scaled to the range [0, 1] using Min-Max normalization. This ensures that features with different units or scales are treated uniformly by the model.

Splitting: The data set is split into training and test sets by engine units to prevent data leaks. An 80-20 split is used for internal validation, ensuring a robust evaluation of model performance.

The preprocessing pipeline is designed to handle the unique characteristics of each subset, allowing the model to generalize across different operating conditions and failure modes.

IV. METHODOLOGY

A. Random Forest Regression

Random Forest is a tree-based ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. This method is particularly effective for highdimensional data sets as it can capture complex, non-linear relationships between features and the target variable.

Key features of Random Forest that make it suitable for RUL prediction include:

Robustness: The model is resistant to noise and overfitting due to its ensemble nature. By averaging predictions from multiple trees, the effects of biased or overfitting individual trees are mitigated.

Feature Importance: Random Forest provides inherent feature importance scores, allowing insights into the relative contribution of each feature to the model. This is especially useful for feature selection and model interpretation.

Parallelization: The training process can be parallelized, making it computationally efficient. This is crucial when working with large data sets, as it allows you to take advantage of multiple processing cores to speed up training.
Hyperparameters:

Number of estimators: 100 decision trees were used, providing a balance between computational efficiency and

predictive performance. A greater number of estimators can improve precision, but also increases computing time.

Maximum Depth: This parameter was adjusted based on the complexity of each data subset, ensuring that the model captures relevant patterns without overfitting the training data.

Random state: A fixed random seed (42) was used to ensure model reproducibility, allowing results to be consistent across each run.

The model was trained separately on each subset of the CMAPSS dataset. Feature selection focused on maintaining all 21 sensors and three operational configurations, as their contributions varied by subset.

B. Explainability Techniques

LIME:

LIME provides localized interpretability by perturbing the input data and fitting a simpler model (e.g., linear regression) to approximate the behavior of the complex model in the vicinity of a specific prediction. This allows users to identify the most influential features on individual predictions. In this study, LIME was applied to randomly selected test instances from each data subset, generating visual explanations that highlight key features driving the model's RUL predictions[6].

SHAP:

SHAP assigns Shapley values to each feature, quantifying its average marginal contribution to the model predictions. This game theory-based approach provides both global and local interpretability, making it ideal for identifying features that are consistently important across different subsets of data. SHAP summary plots and decision plots were used to visualize the influence of features on RUL predictions. Summary plots allow you to see the overall importance of each feature in the model, while decision plots show how variation in features directly influences model predictions.

These explainability techniques allow us to understand not only what predictions the model makes, but also how and why those predictions are generated, increasing confidence in their use in practical applications.

C. Computational Framework: PySpark

PySpark, the Python API for Apache Spark, played a crucial role in our computational infrastructure, enabling efficient large-scale data processing and distributed computing for the Remaining Useful Life (RUL) prediction task. The distributed computing capabilities of PySpark were instrumental in handling the NASA C-MAPSS dataset, which contains complex sensor and operational data across multiple subsets. Key advantages of using PySpark in this study include:

Distributed Data Processing: PySpark's DataFrame API allowed parallel processing of the C-MAPSS dataset, significantly reducing computational time compared to traditional singlemachine approaches. This was particularly beneficial when working with the more complex subsets (FD003 and FD004) that involve multiple operating conditions and failure modes.

Data Preprocessing at Scale: The PySpark DataFrame transformations were used to implement the preprocessing steps outlined earlier, including: RUL calculation, Sensor feature selection, Min-Max normalization, Train-test splitting with engine unit-based separation.

Model Training Optimization: While the Random Forest model was implemented using scikit-learn, PySpark's distributed computing environment facilitated efficient hyperparameter tuning and cross-validation processes.

The implementation leveraged PySpark's MLlib library for distributed machine learning capabilities, complementing the Random Forest regression approach. By utilizing PySpark, we ensured scalability and computational efficiency throughout the data analysis and model training pipeline.

V .RESULTS

A. Model Performance

The performance of the Random Forest model was evaluated on the four subsets using three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2. These metrics quantify the accuracy of the model's predictions and its ability to explain variation in the data.

Table 1. Performance Metrics for PySpark Models

Dataset	Model	MAE	MSE	R^2
FD001	Random Forest	24.72	1121.75	0.3504
FD002	Random Forest	23.32	994.86	0.6560
FD003	Random Forest	31.09	1892.59	-0.1045
FD004	Random Forest	31.97	1900.73	0.3606

The results indicate that the model performs well in the simplest subsets (FD001 and FD002), where the operating conditions are less variable. However, its performance decreases in the most complex subsets (FD003 and FD004), highlighting the need for further optimization to improve performance in more challenging scenarios.

For FD002, the comparative analysis of additional models demonstrates the effectiveness of Gradient Boosting:

Table 2. Comparative Analysis of PySpark and Other Models on FD002

Model	MAE	RMSE	R^2
Decision Tree	30.87	39.78	0.45
Random Forest	28.007	33.24	0.61
Gradient Boosting	24.37	31.91	0.64

As visualized in Fig.1 Gradient Boosting achieved the best performance on FD002, with the lowest MAE and RMSE and the highest R^2 value (0.64). This highlights its potential to handle data sets of moderate complexity effectively. Random Forest followed closely behind, demonstrating balanced performance across all metrics. Decision Tree lagged behind, suggesting limitations in capturing complex patterns in the data.

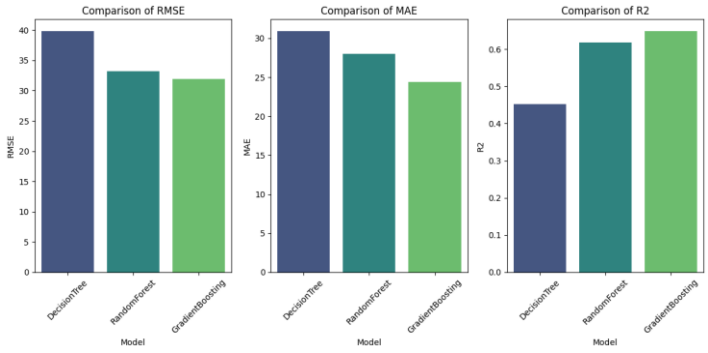

Figure 1

B. Interpretability Analysis

The integration of LIME and SHAP techniques provided critical insight into the model's decision-making process, revealing the specific contributions of sensors and operational configurations to RUL predictions.

LIME results:

LIME analysis showed that predictions for simpler subsets, such as FD001, were primarily influenced by a limited set of sensors (e.g., Sensor 11 and Sensor 15)[9]. These sensors correspond to key temperature and pressure metrics, which are aligned with known degradation patterns in turbofan engines.

For FD002, which involved multiple operating conditions,

LIME highlighted additional features, such as Sensor 3 (related to airflow) and Operating Configuration 1, as significant. This indicates that variations in environmental and operational factors play a more prominent role in influencing RUL under diverse conditions. LIME visualizations revealed that predictions for individual engines often depended on a small subset of features, demonstrating the localized impact of operating conditions on engine health.

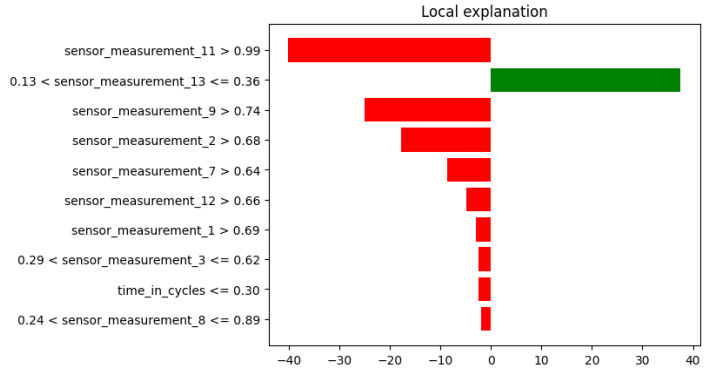

Figure 2

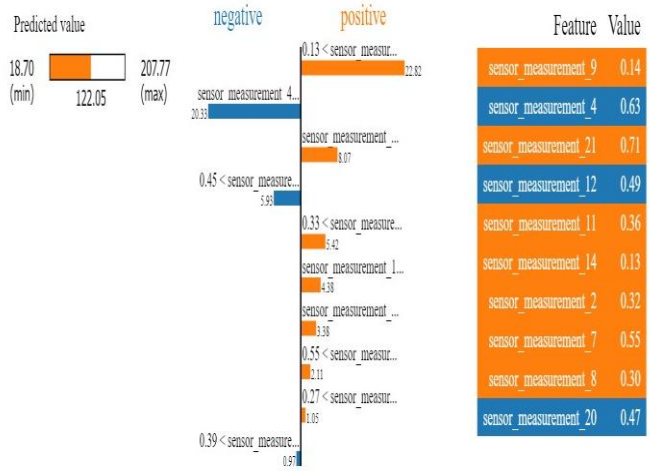

Figure 3

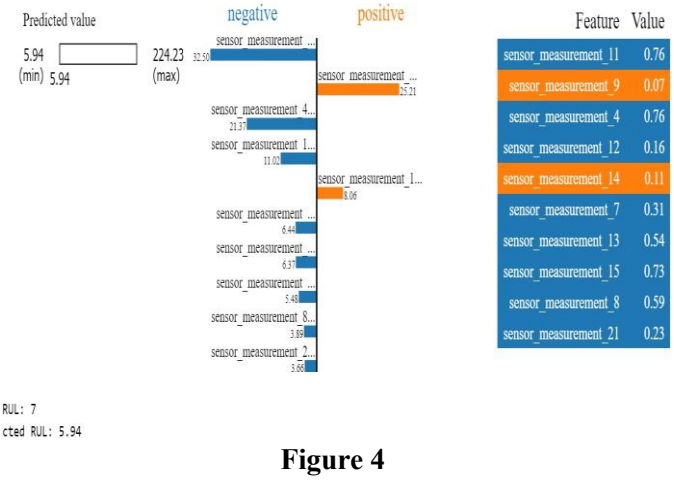

Figure 4

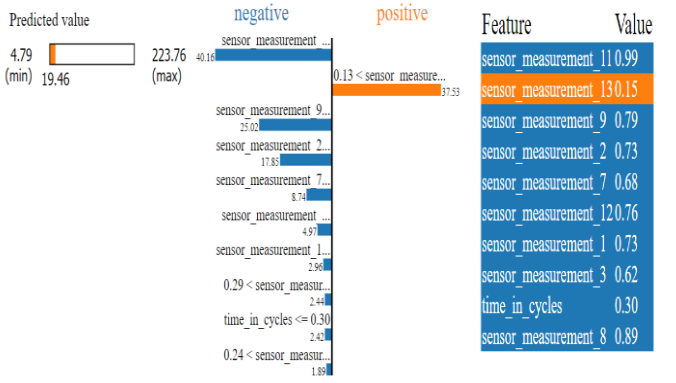

Figure 5

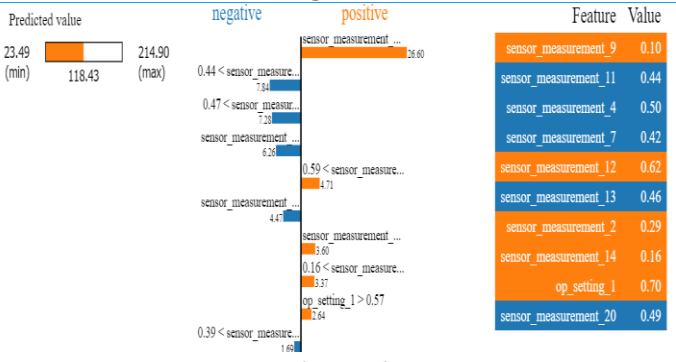

Figure 6

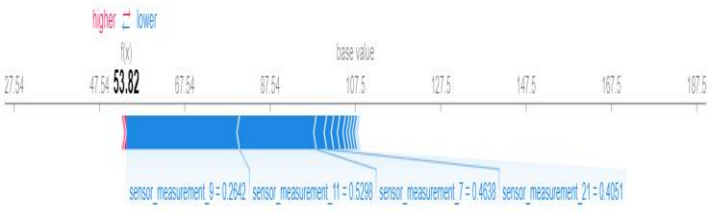

Figure 7

SHAP results:

SHAP summary plots identified Sensor 2 (measuring pressure ratios), Sensor 7 (fuel flow), and Operating Configuration 1 as consistently important features across all subassemblies. These findings underscore the critical role of fuel efficiency and environmental conditions in engine degradation.

SHAP decision graphs provided a comprehensive view of how individual characteristics contributed to specific predictions. For example, engines operating under high pressure ratios showed sharp drops in predicted RUL, emphasizing the impact of operational stress on engine health.

For FD003 and FD004, which involve multiple failure modes, SHAP revealed that the feature contributions were more dispersed, indicating the interaction of multiple degradation mechanisms. This is consistent with the increased complexity of these subassemblies, where failures can be caused by several operating and degradation conditions simultaneously.

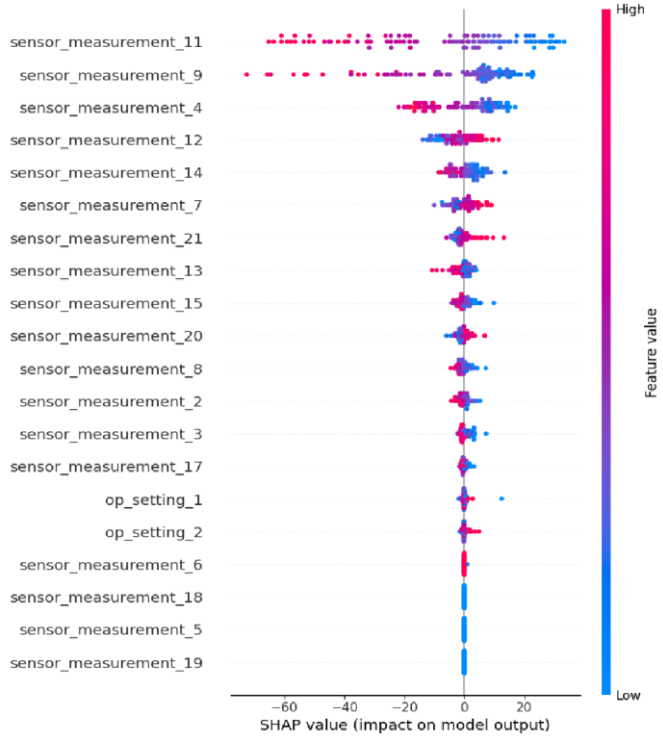

Figure 8

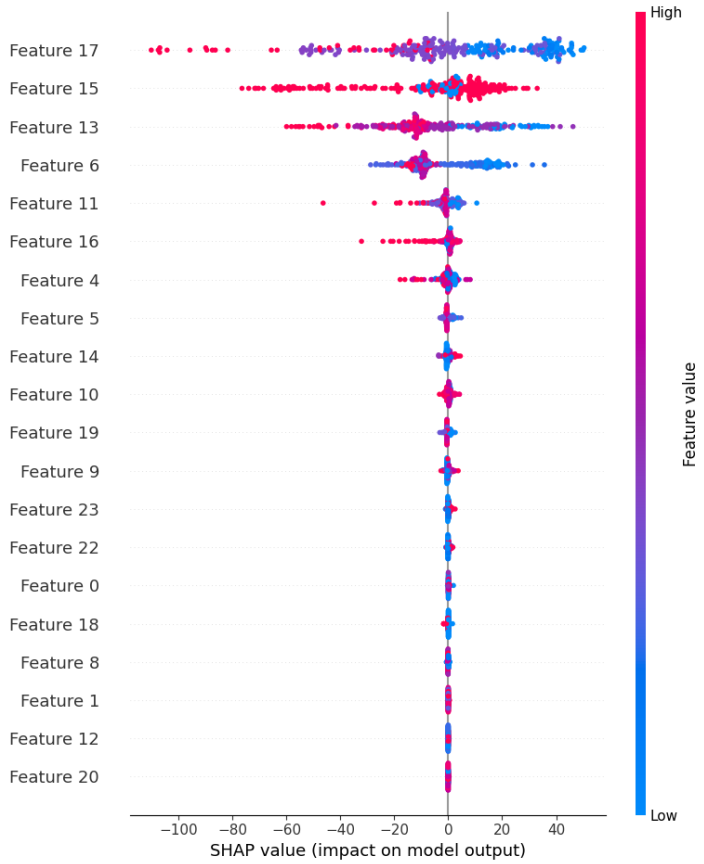

Figure 9

C. Comparison Across Datasets

The results highlight the varied complexity of the four subsets and the model's ability to adapt to different scenarios:

FD001 and FD002: The Random Forest model achieved strong performance, with low MAE values and high R^2 values, particularly in FD002. The consistent operating conditions at FD001 allowed the model to identify clear degradation patterns, while the diverse conditions at FD002 demonstrated the robustness of the model in adapting to variations in operating parameters.

FD003 and FD004: The model faced challenges in these subsets due to the presence of multiple failure modes and greater variability in operating conditions. The negative value of R^2 for FD003 indicates that the model had difficulty capturing the underlying relationships, possibly due to overlapping failure mechanisms, which complicated the identification of consistent patterns of degradation in the engines.

VI. DISCUSSION

The results of this study demonstrate the potential of combining Random Forest regression with explainability techniques for RUL prediction. The following key observations emerged from the analysis:

Predictive Accuracy vs. Complexity:
The model performed well with data subsets with simpler failure modes (FD001 and FD002), but faced challenges with more complex scenarios (FD003 and FD004). This suggests that additional feature engineering or advanced assembly methods may be necessary to improve performance on data sets with high variability. The model's ability to adapt to simpler scenarios is limited when failure mechanisms interact in complex ways, requiring greater specialization in the data.

Importance of Characteristics and Interpretability:
LIME and SHAP provided valuable insights into the features that drive RUL predictions. For example, the continued importance of Sensor 2 and Operational Configuration 1 highlights their critical role in monitoring engine health[3,4]. These findings can inform targeted maintenance strategies, such as prioritizing inspections for engines operating under high-pressure conditions. Understanding which characteristics most affect predictions allows engineers to make more informed decisions about when and where to intervene.

Limitations of Current Approaches:
Although the Random Forest model proved to be robust, its performance on FD003 and FD004 indicates potential limitations in handling complex interactions between failure modes. Furthermore, reliance on predefined features could miss latent patterns that could be captured by deep learning models. However, the lack of interpretability of deep learning models remains a challenge, highlighting the value of the hybrid approach proposed in this study[9]. The ability to obtain clear explanations from Random Forest compensates for the limitations of deep learning methods in terms of transparency.

Real World Applicability:
Integrating explainability techniques ensures that model predictions are actionable and reliable. This is particularly important in aerospace applications, where maintenance decisions must be based on clear evidence. The ability to identify critical sensors and operational factors allows engineers to focus their efforts on priority areas, reducing costs and improving safety[8]. The explainability of the model ensures that decisions made from its predictions are justified and understandable, which is essential for trust in the technology .

VII. CONCLUSION

This study presents a comprehensive framework for RUL prediction using Random Forest regression, supported by LIME and SHAP for interpretability. The model demonstrated strong predictive performance on simpler data sets and provided actionable insights into feature contributions. Key findings include the identification of critical sensors (e.g. Sensor 2, Sensor 11) and operational settings that influence engine degradation[1-10]. These insights can inform maintenance strategies, ensuring timely interventions and minimizing operational disruptions.
However, the model's performance on more complex data subsets highlights the need for further research. Future work will explore advanced assembly techniques, such as Gradient Boosting Machines and hybrid deep learning architectures, to address the challenges posed by multiple failure modes and diverse operating conditions. Furthermore, real-time implementation of the proposed framework will be investigated, enabling dynamic prediction of RUL in live operational scenarios.

Acknowledgments
The authors extend their gratitude to NASA for providing the C-MAPSS data set[6], which serves as a crucial resource for RUL prediction research. They also thank the LIME and SHAP open source contributors for developing tools that improve the interpretability of machine learning models.

References

1. Saxena, A., & Goebel, K. (2008). *C-MAPSS Dataset for Prognostics*.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Why Should I Trust You? Explaining Machine Learning Predictions with LIME*.
3. Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions with SHAP*.
4. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
5. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 173
6. Patrick Seebold, Murali Krishna Kaye, Chang S. Nam *Explainable AI-based Shapley Additive Explanations for Remaining Useful Life Prediction using NASA Turbofan Engine Dataset*

7. L. Chen, L. Wei, Y. Wang, J. Wang, & W. Li, "*Monitoring and predictive maintenance of centrifugal pumps based on smart sensors,*" Sensors, 2022, vol. 22, no. 6, p. 2106, doi: 10.3390/s22062106.

8. D. K. Frederick, J. A. DeCastro, and J. S. Litt, *User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS),* Oct. 2007, Accessed: Dec. 22, 2023.

9. 2014, Anouar BOUROKBA, Ridha EL HAMDI and Mohamed NJAH
A Shapley based XAI approach for a turbofan RUL estimation

10. Y. H. Sheu, (2020). *Illuminating the Black Box: In terpreting Deep Neural Network Models for Psychiatric Research.* Frontiers in Psychiatry.