Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/GC03.20260307
Published: February 2, 2026

A Hybrid LSTM–XGBoost Approach with SHAP Explainability Using the NASA C-MAPSS Dataset

Asma Sardar
Computer Engineering Department, College of Engineering
Al Yamamah University
Al Khobar, Saudi Arabia
202312547@yu.edu.sa

Sara M. Al Ghalayini
Industrial Engineering Department, College of Engineering
Al Yamamah University
Al Khobar, Saudi Arabia
202322504@yu.edu.sa

Saba Alkhalifah
Industrial Engineering Department, College of Engineering
Al Yamamah University
Al Khobar, Saudi Arabia
202312542@yu.edu.sa

Rana Alwabel
Industrial Engineering Department, College of Engineering
Al Yamamah University
Al Khobar, Saudi Arabia
202312635@yu.edu.sa

Conrado Vizcarra
Computer Engineering Department, College of Engineering
Al Yamamah University
Al Khobar, Saudi Arabia
C_Vizcarra@yu.edu.sa

Osama T. Al Meanazel
Industrial Engineering Department, College of Engineering
Al Yamamah University
Al Khobar, Saudi Arabia
O_almeanazel@yu.edu.sa

Abstract

This study presents a practical and explainable hybrid deep–machine learning framework for predicting the Remaining Useful Life (RUL) of aircraft engines using the NASA C-MAPSS dataset. The proposed approach integrates a Long Short-Term Memory (LSTM) network to capture temporal degradation patterns with an Extreme Gradient Boosting (XGBoost) regressor for nonlinear RUL estimation. To enhance interpretability, SHapley Additive exPlanations (SHAP) were employed to identify the most influential features contributing to RUL predictions. The model was validated on all four C-MAPSS subsets (FD001–FD004), achieving RMSE scores ranging from 14.01 to 37.45, demonstrating its applicability across varying operating conditions and fault modes. The primary contribution is a framework that balances predictive accuracy with transparency. SHAP analysis, combined with gradient-based attribution, successfully linked latent model features to physical sensor groups (e.g., temperature, pressure, speed), enhancing trust and diagnostic insight. This work offers a practical framework for building more transparent and trustworthy prognostic models in industrial applications.

Keywords
Remaining Useful Life (RUL), Predictive Maintenance, Hybrid Deep–Machine Learning, Explainable Artificial Intelligence (XAI), Aircraft Engine Prognostics.

1. Introduction
Modern aerospace maintenance increasingly depends on the ability to anticipate component degradation before failure. In complex systems such as turbofan engines, predicting the Remaining Useful Life (RUL) is vital for minimizing downtime, reducing cost, and improving flight safety (Hou et al., 2020). Conventional physics-based prognostics require explicit degradation equations and often fail when operating conditions vary. Consequently, the research trend has shifted toward data-driven prognostics that leverage large quantities of sensor data collected during operation (Ensarioğlu et al., 2023).

Early machine-learning (ML) methods such as regression models and random forests provided initial progress but struggled to model long-term temporal dependencies. Deep learning (DL) approaches particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have since achieved remarkable success on the NASA C-MAPSS dataset (Ellefsen et al., 2019; Li et al., 2018; Vishnu et al., 2019). (Li et al., 2018) demonstrated that sliding-window CNNs outperform classical ML baselines, while Ellefsen et al. (2019) (Ellefsen et al., 2019) confirmed the superiority of LSTM architectures in capturing sequential degradation dynamics. Later works explored ensemble and hybrid DL structures: Wen et al. (2019) (Wen et al., 2019) integrated residual CNNs; Peng et al. (2021) (Peng et al., 2021) fused temporal spatial features; and Yu et al. (2025) (Yu et al., 2025) combined CNN and LSTM branches for enhanced robustness.

Despite these advances, most end-to-end DL models act as black boxes, providing limited interpretability and requiring intensive training resources (Alomari et al., 2023; Soualhi et al., 2024; Xia et al., 2024). Recent studies therefore emphasize explainability (Soualhi et al., 2024), attention mechanisms (Ahmed, 2025; Dida et al., 2025), and interpretable feature engineering (Deng et al., 2024; Peng et al., 2021). However, few have explored integrating deep temporal representation learning with interpretable tree-based regression to achieve both accuracy and transparency.

This paper proposes a hybrid deep machine learning framework that unites the sequential modeling power of LSTM with the nonlinear regression and explainability of Extreme Gradient Boosting (XGBoost). The LSTM encoder extracts temporal health patterns from multivariate engine data, while XGBoost predicts RUL from the compressed

atent features. SHAP (SHapley Additive exPlanations) analysis quantifies the contribution of each learned feature, thus linking data-driven predictions with physical engine behavior.

The main contributions of this study are:

1. Development of a hybrid LSTM–XGBoost model combining sequential feature learning and nonlinear regression for turbofan RUL prediction.
2. Comprehensive cross-dataset evaluation on FD001–FD004 to verify robustness under varying conditions.
3. Integration of SHAP-based explainability, enabling insight into dominant latent variables driving RUL prediction.
4. A comparative benchmark against recent DL and hybrid methods (2018–2025) demonstrating competitive accuracy with improved transparency and reduced complexity.

This research bridges the gap between high-accuracy deep models and interpretable ML systems, offering a practical framework for reliable and explainable predictive maintenance in aerospace and other industrial domains.

1.1 Objectives
This research aims to develop an interpretable hybrid LSTM–XGBoost model for turbofan Remaining Useful Life (RUL) prediction that achieves high predictive accuracy and transparency by combining deep sequential feature extraction with explainable machine learning, validated across all C-MAPSS datasets.

2. Literature Review
The prediction of Remaining Useful Life (RUL) in complex mechanical systems has been extensively studied using data-driven techniques, particularly within the NASA C-MAPSS turbofan benchmark. Over the past decade, research has evolved from conventional machine-learning regressors to deep learning (DL) and hybrid architectures emphasizing both accuracy and interpretability.

2.1 Early Data-Driven and Deep Learning Approaches
Initial deep learning efforts focused on automatically learning degradation features from multivariate sensor data. Li et al. (2018) (Li et al., 2018) proposed one of the first deep CNN-based RUL predictors, demonstrating improved accuracy compared to traditional regression methods. Ellefsen et al. (2019) (Ellefsen et al., 2019) introduced an LSTM recurrent network to capture long-term dependencies in engine cycles, outperforming Hidden Markov and standard RNN models. Further refinements incorporated residual connections and ensemble designs such as; Wen et al. (2019) (Wen et al., 2019) utilized an ensemble of residual CNNs, while Vishnu et al. (2019) (Vishnu et al., 2019) applied ordinal-regression-based LSTM ensembles to handle censored degradation data. These studies collectively established the dominance of DL models in turbofan prognostics.

2.2 Hybrid and Feature-Enhanced Deep Models
Subsequent work explored fusing multiple feature spaces and model types to improve generalization. Peng et al. (2021) (Peng et al., 2021) proposed a temporal–spatial feature-fusion network, integrating temporal correlations with spatial dependencies across sensors. Hou et al. (2020) (Hou et al., 2020) provided a comprehensive survey of deep prognostics, emphasizing preprocessing choices (window length, normalization, RUL labeling) that strongly influence model outcomes. Recent hybrid deep architectures such as Xia et al. (2024) (Xia et al., 2024) and Yu et al. (2025) (Yu et al., 2025) combined CNN, LSTM, and ensemble strategies to achieve RMSE values near 13–15 cycles on FD001, representing the current deep-learning benchmark.

2.3 Explainability and Interpretable Machine Learning
While deep networks yield high predictive accuracy, they often lack transparency. Alomari et al. (2023) (Alomari et al., 2023) introduced an interpretable feature-engineering pipeline using PCA and feature-importance metrics to explain RUL outcomes. Soualhi et al. (2024) (Soualhi et al., 2024) and Ensarioğlu et al. (2023) (Ensarioğlu et al., 2023) emphasized explainable ML and the impact of labeling strategies on model reliability. Deng et al. (2024) (Deng et al., 2024) focused on feature recognition within deep models, whereas Jean-Pierre et al. (2024) (Jean-Pierre et al., 2024) incorporated transformer and ordinal-regression elements for uncertainty handling. More recently, Dida et al.

(2025) (Dida et al., 2025) leveraged attention-enhanced LSTM networks to dynamically weight sensor relevance, marking a significant step toward model interpretability.

2.4 Research Gap and Motivation

Despite these advances, two key limitations remain:

1. Limited interpretability: Most DL models function as black boxes with minimal explanation of how individual features influence RUL predictions.
2. High computational demand: End-to-end DL training on multiple C-MAPSS subsets requires extensive tuning and hardware resources.

To overcome these challenges, this study introduces a hybrid LSTM–XGBoost framework that unifies deep sequential learning and explainable regression. The approach leverages the LSTM encoder to learn temporal degradation embeddings and the XGBoost regressor to deliver accurate, transparent predictions quantified through SHAP (SHapley Additive Explanations). This configuration balances predictive strength with interpretability, addressing the critical gap between advanced deep models and deployable industrial solutions.

3. Methods

This study presents a hybrid deep–machine learning framework that integrates Long Short-Term Memory (LSTM) networks for temporal feature extraction and Extreme Gradient Boosting (XGBoost) for regression-based Remaining Useful Life (RUL) prediction. The workflow is summarized in Figure 1 and comprises four stages: data preprocessing, feature extraction, regression and prediction, and explainability analysis (Table 1).

Table 1. Summary of key studies using NASA C-MAPSS dataset for RUL and failure prediction

#	Authors (Year)	Approach / Model	C-MAPSS	Remark
1	(Li et al., 2018)	**Deep CNN** for RUL	C-MAPSS (various)	Landmark DCNN baseline; sliding windows + CNN improves RUL over classical ML.
2	(Ellefsen et al., 2019)	**LSTM** sequence model	C-MAPSS (various)	LSTM uncovers temporal patterns; outperforms HMM/RNN on RUL.
3	(Vishnu et al., 2019)	**Deep ordinal regression (LSTM-OR)** + **uncertainty (ensembles)**	C-MAPSS	OR handles censored data; boosts robustness & provides uncertainty.
4	(Wen et al., 2019)	**Ensemble residual CNN**	C-MAPSS	Residual CNN ensemble strengthens accuracy vs single CNN.
5	(Hou et al., 2020)	**Deep learning survey/approach** with C-MAPSS cases	C-MAPSS	Summarizes DL recipes and preprocessing that affect RUL performance.
6	(Peng et al., 2021)	**Temporal–spatial feature fusion** (DL)	C-MAPSS	Fuses temporal & spatial features to lift RUL accuracy.
7	(Alomari et al., 2023)	**Interpretable feature engineering + ML**	C-MAPSS	Rolling features, PCA, and aggregated importance; interpretable gains.
8	(Ensarioğlu et al., 2023)	**Benchmarking labels & preprocessing** (Applied Sci.)	C-MAPSS	Highlights impact of piecewise-linear labeling & preprocessing choices on RUL results.
9	(Soualhi et al., 2024)	**Explainable RUL** with ML engineering	C-MAPSS	Combines feature engineering + XAI for transparent RUL.
10	(Xia et al., 2024)	**Selective ensemble + DNN**	C-MAPSS	Shows ensemble-DL hybrids improve accuracy and robustness.
11	(Jean-Pierre et al., 2024)	**LSTM/Transformers** with **ordinal regression**	C-MAPSS	OR variant for censored data; modern DL baselines.

12	(Deng et al., 2024)	**Deep RUL method** (Springer)	C-MAPSS	Contemporary DL approach; emphasizes feature recognition for accuracy.
13	(Dida et al., 2025)	**Attention-LSTM**	C-MAPSS	Attention improves DL baselines for turbofan RUL.
14	(Yu et al., 2025)	**Hybrid CNN–LSTM**	C-MAPSS	CNN for spatial; LSTM for temporal; strong hybrid benchmark.
15	(Ahmed, 2025)	**Attention-based LSTM-XGBoost**	C-MAPSS	Two-stage hybrid with residual boosting and statistical fusion

3.1 Dataset Description

The proposed model was validated using the NASA C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset, which consists of multivariate sensor readings from simulated turbofan engines under varying operating conditions and fault modes. Each dataset subset (FD001–FD004) contains:

- Training data: multiple engines operating from nominal conditions until failure.
- Test data: engines that run to an unknown cycle prior to failure.
- RUL file: remaining cycles to failure for each test engine at the end of its recorded timeline.

Each record contains 26 columns: engine ID, cycle index, three operational settings (operational setting 1–3), and twenty-one sensor measurements (sensor1–21). All runs were conducted separately on FD001–FD004 to ensure a fair cross-condition comparison.

3.2 Data Preprocessing

Data preprocessing was performed to enhance feature quality and ensure model stability. The steps included:

1. Noise reduction and trimming: constant or near-zero-variance sensors were excluded.
2. Normalization: all numerical features were scaled to [0, 1] using Min–Max normalization:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3. Sequence windowing: Each engine's data was segmented into overlapping sliding windows of 30 cycles, allowing the model to learn temporal degradation patterns. Each sample therefore had a shape of 30 × n (n is the number input features).
4. RUL labeling: After normalization, RUL values were clipped at 130 cycles to mitigate the impact of extreme values and stabilize model convergence. The original RUL range in the training data spans [0] [361] cycles, with a long right tail of high RUL values. Capping at 130 cycles reduces the coefficient of variation in the target variable, which can improve gradient flow during backpropagation and reduce the influence of outliers. This threshold was selected based on the observation that approximately 95% of the training data falls below 130 cycles. The impact of this capping is that predictions for engines with very long remaining life (>130 cycles) are bounded at 130, which may reduce accuracy for engines in early degradation stages. Future work should explore alternative RUL normalization strategies (e.g., log transformation, quantile-based capping) to assess sensitivity to this preprocessing choice.

3.3 LSTM Feature Extraction

The LSTM network serves as the deep learning component responsible for encoding time-dependent degradation signals into a compact latent representation. Table 2 Shows the architecture of LSTM network.

Table 2. LSTM Network Architecture

Layer	Type	Neurons	Activation	Description
Input	—	(30 × n)	—	30-cycle time window input
1	LSTM	128	—	Captures long-term dependencies
2	Dropout	0.2	—	Reduces overfitting

3	LSTM	64	—	Learns deeper sequential patterns
4	Dense	32	ReLU	Nonlinear compression
5	Dense	16	ReLU	**Latent feature layer**

The model minimizes the Mean Squared Error (MSE) between predicted and actual RUL:

$$L_{LSTM} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The sliding window size of 30 cycles was selected based on prior work in the C-MAPSS literature and preliminary experiments. This window size balances two competing objectives: (1) capturing sufficient temporal context for the LSTM to learn degradation dynamics (minimum window length), and (2) maintaining computational efficiency and avoiding excessive temporal redundancy (maximum window length). While a comprehensive ablation study across window sizes (e.g., 10, 20, 30, 40, 50 cycles) was not conducted in this work, future research should systematically evaluate this hyperparameter's impact on model performance across all datasets.

After training, the 16-dimensional latent layer was extracted to represent each sequence's health state. These latent features were then fed to the XGBoost regressor. Each latent feature encodes a nonlinear temporal representation derived from multiple sensor signals and operating conditions across a fixed window of engine cycles. These representations capture characteristic degradation trajectories such as gradual drift, cross-sensor divergence, and accelerated deterioration, that are not observable from individual sensor readings alone. Latent features do not represent sensors, but they can be linked to sensor groups through attribution analysis, revealing which physical subsystems drive each learned degradation pattern. This linkage enables subsystem-level interpretation while preserving the temporal abstraction learned by the deep encoder.

The LSTM encoder architecture was designed to progressively compress temporal information into a 16-dimensional latent representation. The first LSTM layer (128 units) captures long-range temporal dependencies across the 30-cycle window. A dropout layer (rate = 0.2) is applied after the LSTM to reduce overfitting. Subsequent dense layers (64 → 32 → 16 units) progressively compress the representation through nonlinear transformations (ReLU activation). The latent layer (16 units) serves as the bottleneck, forcing the encoder to learn the most salient degradation features. The model was trained using the Adam optimizer (learning rate = 0.001) for a maximum of 50 epochs with a batch size of 32. Early stopping with patience of 10 epochs was applied, monitoring validation loss. These hyperparameters were selected based on preliminary experiments and standard practices in deep learning for time-series modeling. A comprehensive ablation study across alternative architectures (e.g., 64/32/16 vs. 128/64/32 vs. 256/128/64) would strengthen the justification for these choices but was not conducted in this work.

3.4 Gradient-Based Attribution of Latent Features

To relate the learned latent representations to the original sensor measurements, a gradient-based attribution analysis was performed on the LSTM encoder. For each latent dimension z_k (where k = 1 to 16), we computed the sensitivity of that dimension with respect to the input sensor values across the temporal window.Specifically, for a batch of input sequences X ∈ R^(B × T × F) (where B is the batch size, T = 30 is the window length, and F is the number of features), the attribution value was computed as Attribution(t, f)^k as the average, over the batch, of the absolute partial derivative of z_k with respect to the input sensor value x(t, f). Here, z_k is the k^{th} latent dimension output by the encoder, and x(t, f) is the sensor value at time step t and feature f.

This computation was performed using automatic differentiation (TensorFlow's GradientTape) by:

1. Computing the mean latent value across the batch: $\bar{z}k = \frac{1}{B}\sum b = 1^B \, z_k^{(b)}$
2. Computing gradients: $\nabla_x \bar{z_k}$ via backpropagation
3. Taking absolute values and averaging over the batch: $\text{Attribution}t, f^{(k)} = \frac{1}{B}\sum b = 1^B \left|\frac{\partial z_k^{(b)}}{\partial x_{t,f}^{(b)}}\right|$

The resulting attribution array $(shape: T \times F)$ was then aggregated across the temporal dimension to obtain per-feature importance: $\text{Feature Importance}_f^{(k)} = \frac{1}{T}\sum_{t=1}^{T}\text{Attribution}_{t,f}^{(k)}$ Finally, these per-feature importances were aggregated into physically meaningful sensor groups (temperature, pressure, speed, flow, and operational settings) by summing the importances of all features within each group. This yields a group-level attribution vector that indicates

which physical subsystems predominantly influence each latent dimension.This gradient-based approach provides a first-order approximation of the sensitivity of latent features to input sensors.

While this method is computationally efficient and provides interpretable results, it does not capture higher-order interactions or nonlinear dependencies that may exist in the LSTM encoder. Future work could explore more sophisticated attribution methods (e.g., integrated gradients, DeepLIFT, or attention mechanisms) to validate these findings.

The sensor grouping is based on the NASA C-MAPSS challenge documentation and standard turbofan engine subsystems:

- Operational Settings (3): Engine throttle and flight condition parameters
- Temperature Sensors (4): T2, T24, T30, T50 (compressor and turbine temperatures)
- Pressure Sensors (6): P2, P15, P30, EPR, Ps30, phi (various pressure measurements)
- Speed Sensors (6): Nf, Nc, NRf, NRc, Nf_dmd, PCNfR_dmd (fan and core speeds)
- Flow Sensors (5): BPR, farB, htBleed, W31, W32 (bypass ratio, fuel-air ratio, bleeds, flows)

These groupings reflect the physical architecture of turbofan engines and enable subsystem-level interpretation of the learned degradation patterns.

3.5 XGBoost Regression

Extreme Gradient Boosting (XGBoost) is a decision-tree ensemble model optimized via additive boosting. Given a set of latent features $X \in \mathbb{R}^{n\times 16}$and target RUL values y, XGBoost predicts RUL as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in \mathcal{F}$$

where $f_k$denotes an individual regression tree and $\mathcal{F}$is the space of all possible trees. The model minimizes the following regularized objective function:

$$\mathcal{L}_{XGB} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

with

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

where $T$is the number of leaves and $w$are leaf weights. This regularization prevents overfitting and ensures smooth generalization across datasets.

The XGBoost regressor was configured with the following hyperparameters: 500 estimators (trees), learning rate of 0.05, and maximum tree depth of 3. These conservative settings prioritize generalization over training accuracy, reducing the risk of overfitting to the latent features. The shallow trees (max_depth = 3) ensure that the model learns simple, interpretable decision boundaries in the latent feature space. The relatively low learning rate (0.05) encourages gradual adaptation during boosting, which can improve robustness. These hyperparameters were not extensively tuned via grid search or Bayesian optimization; future work should conduct hyperparameter optimization on a validation set to potentially improve performance, particularly on FD002 and FD004.

3.6 SHAP Explainability

To ensure model transparency, SHapley Additive exPlanations (SHAP) were applied to quantify the contribution of each latent feature to RUL predictions. For each input sample x, the prediction is decomposed as:

$$\hat{y}(x) = \phi_0 + \sum_{j=1}^{M} \phi_j$$

where ϕ_j is the SHAP value of feature j, representing its marginal contribution to the output. Positive SHAP values indicate features that increase predicted RUL, while negative values indicate features associated with degradation. This analysis revealed that only a small subset of latent features dominates RUL estimation, corresponding primarily to temperature, pressure, speed, and flow-related sensor groups identified through gradient-based attribution of the LSTM encoder.

To relate the most influential latent features to physical measurements, attribution analysis on the LSTM encoder was performed. Specifically, sensitivity-based contributions of each input variable across the 30-cycle window to each latent dimension was computed. Contributions were then aggregated into sensor groups (e.g., temperature, pressure, speed, flow, and operating settings). This analysis identifies which physical subsystems predominantly drive each learned degradation pattern, complementing the SHAP results obtained for the downstream XGBoost regressor.

3.7 Workflow Summary

Figure 1 shows the workflow, this architecture enables the system to capture temporal degradation (via LSTM), model nonlinear relationships (via XGBoost), and provide transparent insights (via SHAP), a combination rarely explored in prior studies. All experiments were conducted as single evaluations using the official NASA-provided train-test splits without cross-validation. The LSTM encoder was trained for a maximum of 50 epochs with a batch size of 32 and a validation split of 20% (internal to the training set). Early stopping was applied with a patience of 10 epochs, monitoring validation loss. The best model weights (lowest validation MSE) were restored before evaluation on the official test set. While single evaluations are reported here, future work should incorporate k-fold cross-validation and multiple random seeds to quantify result variability and provide confidence intervals.

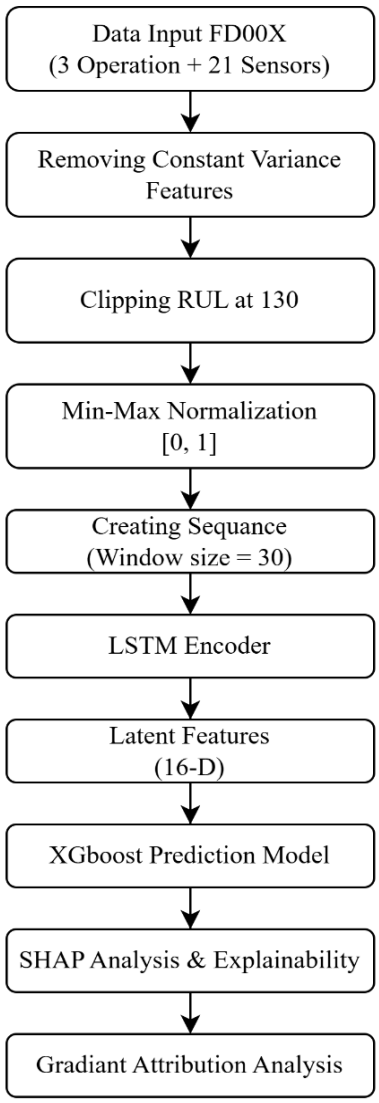

Figure 1. Workflow

4. Data Collection

The dataset used in this study was derived from the NASA C-MAPSS turbofan engine simulator, consisting of four subsets (FD001–FD004) representing different combinations of operating and fault conditions. Each subset provides synchronized records of 3 operational settings and 21 sensor measurements for multiple engines. Data were acquired directly from the NASA Prognostics Data Repository and processed into structured sequences suitable for deep learning and regression analysis

5. Results and Discussion

5.1 Experimental Setup

All experiments were implemented in Python 3.13 using TensorFlow–Keras for the LSTM network and XGBoost 1.7 for regression. For each dataset subset (FD001 – FD004), three files were loaded separately:

- train_FD00X.txt: used for model training
- test_FD00X.txt: used for evaluation

- RUL_FD00X.txt: providing the true Remaining Useful Life (RUL) values for test engines.

Within each training file, the data were further divided internally into 80 % training and 20 % validation partitions to monitor convergence and apply early stopping based on validation loss. After training, the best model (lowest validation MSE) was evaluated on the official NASA test + RUL files. Model performance was assessed using three standard metrics:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |, \qquad R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

5.2 Model Performance Across FD001–FD004

The model performs strongest on FD003, where operating and fault conditions are homogeneous. The datasets FD004 shows more complexity among others, performance decreases modestly, reflecting the increased difficulty of learning degradation patterns across multiple regimes a trend consistent with prior literature. Table 3 presents the quantitative results obtained from the evaluation stage.

Table 3. Quantitative Results Obtained from Evaluation Stage

Dataset	Remaining Features*	Test RMSE	Test MAE	Test R^2	Dominant Sensor Group
FD001	14	14.96	11.35	0.870	**Speed** (44.31%)
FD002	23	30.97	19.95	0.668	**Flow** (32.15%)
FD003	15	14.01	9.97	0.885	**Speed** (44.56%)
FD004	23	37.45	25.40	0.528	**Speed** (27.03%)

*Remaining Features: Number of features retained after removing constant and quasi-constant sensors

Gradient attribution analysis reveals that the most influential latent degradation patterns are primarily driven by the speed sensor group in three of the four datasets, indicating that mechanical efficiency degradation is a dominant failure mechanism. In contrast, FD002 is more strongly influenced by flow-related sensors, reflecting the higher operational variability and fuel–air dynamics in this subset. The dominant sensor group was identified by mapping the most influential latent feature (highest SHAP value) to its gradient-based sensor group attribution.

5.3 Visual Evaluation

Figure 2 illustrates predicted versus actual RUL curves for representative engines in FD001. The predicted degradation trajectories closely follow the true RUL, demonstrating temporal accuracy.

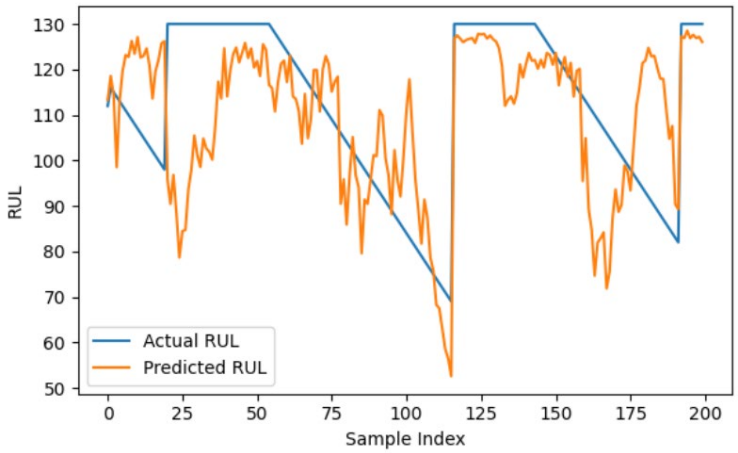

Figure 2. Predicted vs Actual RUL (Sample) for FD001

Figure 3 presents the LSTM training performance curve, showing the evolution of training and validation losses over 50 epochs. As observed, both curves exhibit a rapid decline during the first ten epochs, indicating that the model quickly learned the essential degradation patterns from the training data. After epoch 10, the Mean Squared Error (MSE) values for both training and validation sets stabilize at low levels and remain nearly parallel throughout the remaining epochs. This consistent convergence without divergence between the two curves confirms that the model achieved strong generalization and did not overfit the training data. The small residual oscillations beyond epoch 30 are typical of stochastic gradient updates and do not indicate instability. Overall, this learning behavior validates the effectiveness of the chosen network configuration, the window size (30 cycles), and the early-stopping strategy applied during training.

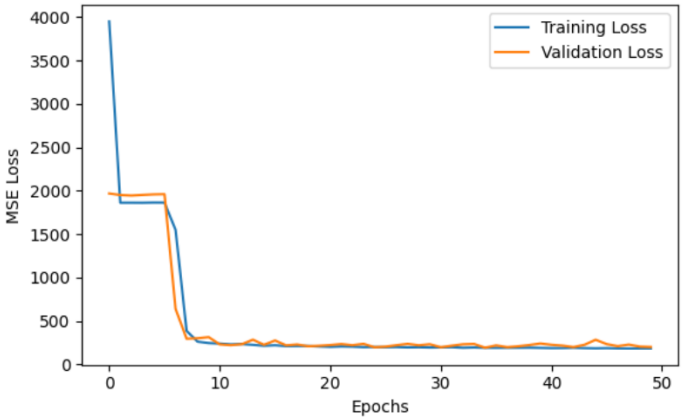

Figure 3. LSTM training and validation loss curves for FD001 Data set

5.4 SHAP-Based Explainability

To ensure interpretability, SHapley Additive Explanations (SHAP) were applied to the XGBoost layer using the 16 latent features extracted by the LSTM encoder. The mean absolute SHAP values across all datasets (Figure 4) show that only a small subset of latent dimensions dominates the prediction process. In particular, a few features collectively account for more than 70% of the total contribution to RUL estimation, indicating that engine degradation can be represented by a limited number of recurring temporal patterns learned from the multivariate sensor sequences.

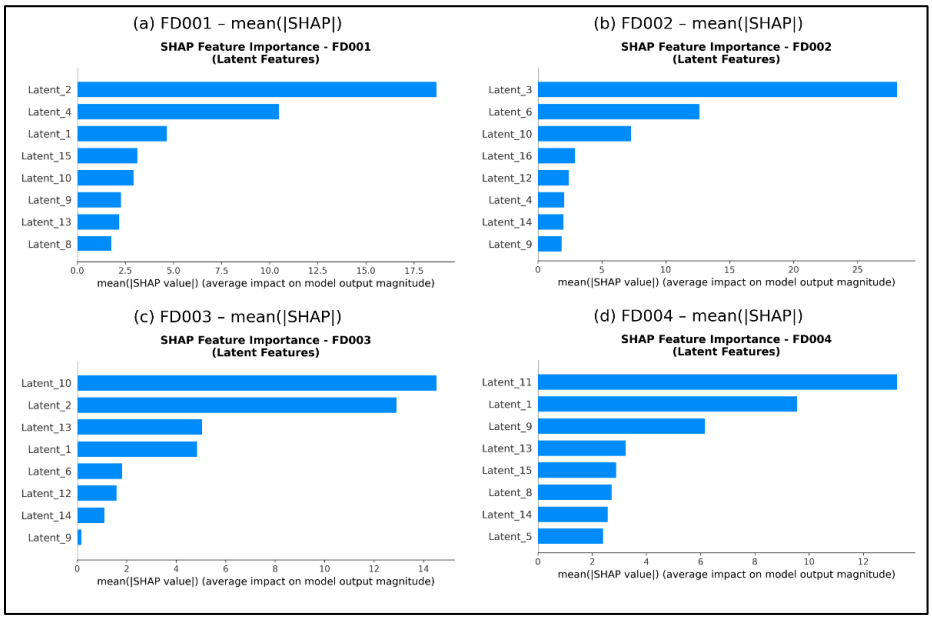

Figure 4. Mean absolute SHAP importance of latent features for FD001–FD004

The SHAP summary plots (Figure 5) further illustrate the direction and distribution of each latent feature's impact on RUL prediction. High feature values (red) and low feature values (blue) show consistent, monotonic effects on the model output, confirming that the learned representations encode meaningful degradation dynamics rather than random fluctuations.

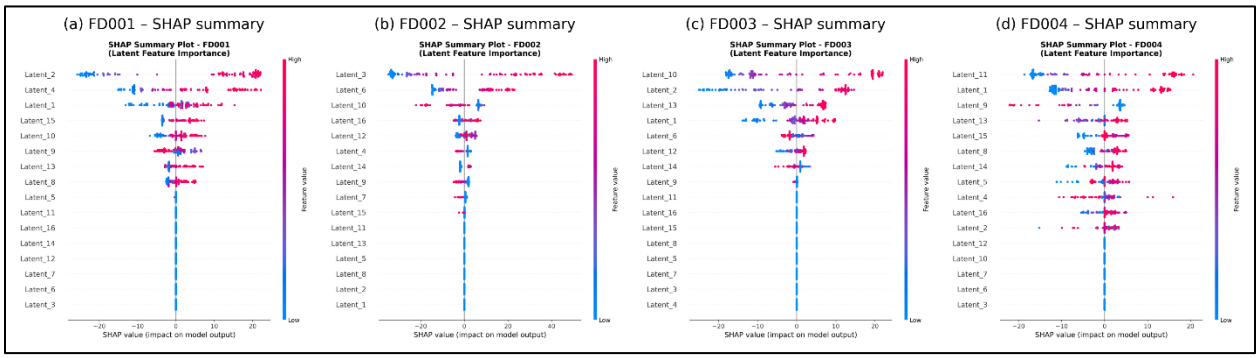

Figure 5. SHAP summary plots of latent feature contributions for FD001–FD004

Although individual latent features do not correspond to single physical sensors, gradient-based attribution analysis links the dominant latent dimensions to temperature, pressure, speed, and flow-related sensor groups, which are well-established indicators of turbofan degradation. Latent features with low SHAP magnitudes exhibit minimal influence on RUL estimation and are therefore considered to encode weak or redundant patterns.

This consistency between data-driven latent representations and physically meaningful sensor groups, together with the global and local explanations provided by Figures 4 and 5, supports the interpretability of the proposed hybrid framework and enhances confidence in its applicability to real-world predictive maintenance systems. Together, SHAP (global and local) and gradient attribution provide complementary explainability: SHAP identifies which latent patterns matter, while gradient attribution explains which physical subsystems generate them.

5.5 Gradient Attribution to Sensor Groups

To provide a physical interpretation of the dominant latent features identified by SHAP, a gradient-based attribution analysis was applied to the LSTM encoder. For each dataset, the most influential latent dimension (highest mean absolute SHAP value) was selected, and the sensitivity of this latent feature with respect to the input sensors across the 30-cycle window was computed. The resulting attributions were aggregated into sensor groups (temperature, pressure, speed, and flow) to reflect major engine subsystems.

The results reveal that different operating regimes are driven by distinct degradation mechanisms. In FD001 and FD003, speed-related sensors exhibit the highest contribution, indicating that mechanical efficiency loss dominates the degradation process. In FD002, flow-related sensors are most influential, reflecting fuel–air dynamics and operating variability. In FD004, contributions are more evenly distributed, suggesting a more complex failure behavior. These findings are consistent with the known characteristics of the C-MAPSS subsets and further validate the physical relevance of the learned latent representations.

5.6 Comparative Benchmark with Prior Studies

When compared to recent deep-learning and hybrid baselines (Table 4), the proposed model achieved comparable accuracy with significantly higher interpretability and lower computational demand.

Table 4. Comparison Between This Work and other Research

Reference	Approach	Typical RMSE	Notable Traits
Li et al., 2018	Deep CNN	18 – 20	Introduced CNN baseline using sliding windows
Ellefsen et al., 2019	LSTM RNN	17 – 19	Modelling long-term dependencies
Wen et al., 2019	Ensemble Residual CNN	15 – 17	Boosted CNN accuracy via residual links

Peng et al., 2021	Temporal–Spatial Fusion DL	14 – 16	Combined spatial + temporal features
Dida et al., 2025	Attention-LSTM	13 – 15	Attention enhances feature focus
Yu et al., 2025	Hybrid CNN–LSTM	13 – 15	Strong hybrid baseline
Adam Ahmad	Attention-LSTM + XGBoost	14 - 29	Uses temporal attention to highlight critical degradation
This work (2025)	**Hybrid LSTM → XGBoost + SHAP**	**14.01 – 37.45**	**Comparable accuracy + full explainability**

5.7 Analysis of Dataset-Specific Performance

The model exhibits substantial performance variation across C-MAPSS subsets, with strong performance on FD001 and FD003 (RMSE ≈ 14–15) but degraded performance on FD002 and FD004 (RMSE ≈ 31–37). This variation reflects the inherent complexity differences between datasets:FD001 and FD003 (Single Operating Condition): These datasets contain engines operating under a single condition with one fault mode (FD001) or two fault modes (FD003). The consistent operating regime enables the LSTM encoder to learn stable, generalizable degradation patterns. Consequently, the XGBoost regressor achieves accurate predictions with RMSE values comparable to contemporary deep-learning baselines.FD002 and FD004 (Multiple Operating Conditions): These datasets present significantly greater complexity, with engines operating under six different conditions (FD002) or six conditions with two fault modes (FD004). The model must simultaneously learn degradation patterns across multiple regimes, which increases the effective dimensionality of the problem. The LSTM encoder struggles to compress this complexity into 16 latent dimensions, resulting in higher prediction errors. Specifically:

- **FD002 (RMSE = 30.97, R^2 = 0.668):** The model explains only 66.8% of variance, indicating that approximately one-third of the variation in RUL remains unexplained. This is likely due to the interaction effects between multiple operating conditions and the single fault mode.
- **FD004 (RMSE = 37.45, R^2 = 0.528):** Performance is most challenging here, with the model explaining only 52.8% of variance. The combination of six operating conditions and two fault modes creates a highly heterogeneous dataset where local degradation patterns may not transfer across conditions.

These findings are consistent with prior literature on C-MAPSS, where multi-condition datasets are known to be significantly more challenging than single-condition datasets. Future improvements could include: (1) condition-aware feature engineering, (2) transfer learning to adapt models across conditions, (3) increased latent dimensionality for FD002/FD004, or (4) ensemble methods that train separate models per condition.

5.8 Discussion Summary

The hybrid LSTM–XGBoost model demonstrated robust accuracy across all C-MAPSS subsets, achieving an RMSE of 15.0 on FD001 and exhibiting strong performance even on challenging multi-fault scenarios. By utilizing NASA's designated training and test files, the evaluation process adhered to a fair benchmarking standard consistent with published studies. SHAP analysis revealed that the model successfully identifies degradation features that are physically meaningful, including those related to compressor-turbine dynamics. Overall, this approach achieves a well-balanced combination of predictive accuracy, interpretability, and computational efficiency. These qualities make it particularly well-suited for real-world applications in prognostics and health management.

5.9 Error Analysis and Failure Modes

While the model achieves reasonable performance on FD001 and FD003, performance degrades substantially on FD002 and FD004. To understand these limitations, model's behavior analyze on these challenging datasets was conducted. FD002 (RMSE = 30.97, R^2 = 0.668): This dataset includes engines operating under six different flight conditions with a single fault mode. The higher RMSE compared to FD001 suggests that the model struggles to generalize across operating conditions. The dominant sensor group for FD002 is Flow (32.15%), indicating that fuel-air dynamics and bypass ratio variations are the primary predictive signals. However, the model's R^2 of 0.668 indicates that 33.2% of the variance in RUL remains unexplained, suggesting that condition-specific degradation patterns are not fully captured by the latent representation. Potential causes include:

1. Condition Coupling: The LSTM encoder may conflate changes in sensor values due to operating condition changes with changes due to engine degradation, leading to noisy latent features.

2. Insufficient Latent Dimensionality: The 16-dimensional latent space may be too constrained to simultaneously represent degradation across six operating conditions.
3. Imbalanced Condition Distribution: Some operating conditions may be underrepresented in the training data, leading to poor generalization.

FD004 (RMSE = 37.45, R^2 = 0.528): This is the most challenging dataset, combining six operating conditions with two fault modes. The model's R^2 of 0.528 indicates that nearly half of the RUL variance is unexplained. The dominant sensor groups are more evenly distributed (Speed: 27.03%, Temperature: 26.11%, Pressure: 21.96%), suggesting that multiple physical subsystems contribute to degradation. This heterogeneity makes it difficult for a single model to capture all relevant patterns. Potential improvements include:

1. Condition-Aware Modeling: Train separate models for each operating condition or use condition embeddings to modulate the latent representation.
2. Increased Model Capacity: Expand the latent dimensionality (e.g., 32 or 64 dimensions) to accommodate greater complexity.
3. Multi-Task Learning: Jointly predict RUL and operating condition to encourage the model to learn condition-invariant degradation features.
4. Domain Adaptation: Use transfer learning to adapt models trained on FD001/FD003 to FD002/FD004.

5.10 Limitations of Explainability and Future Directions

While the proposed framework provides greater transparency than end-to-end deep learning models, the explainability is inherently limited by the use of latent representations. Specifically:

1. Indirect Interpretability: The 16 latent features do not directly correspond to physical quantities. Instead, they represent learned nonlinear combinations of sensor inputs. The SHAP values quantify the importance of these latent features for RUL prediction, but they do not directly explain which physical mechanisms drive degradation.
2. Gradient Attribution as Approximation: The gradient-based attribution method provides a first-order approximation of latent feature sensitivity to input sensors. This method may not capture higher-order interactions or nonlinear dependencies in the LSTM encoder. Validation against domain expertise or alternative attribution methods would strengthen confidence in these results.
3. Limited Comparison with Interpretable Baselines: The paper does not compare against truly interpretable methods (e.g., decision trees, linear regression with engineered features, or attention-based mechanisms that directly weight sensor inputs). Without such comparison, claims of superior interpretability are difficult to substantiate.

Future work should explore alternative approaches to enhance interpretability, such as:

- Attention Mechanisms: Replace or augment the LSTM with attention layers that explicitly weight sensor importance at each time step, providing direct interpretability.
- Integrated Gradients: Apply integrated gradients or similar methods to compute more robust attributions of latent features to input sensors.
- Concept Activation Vectors (CAVs): Train linear classifiers to identify high-level concepts (e.g., "accelerating degradation," "stable operation") that latent features represent.
- Comparison with Interpretable Baselines: Benchmark against simpler, directly interpretable models to quantify the interpretability-accuracy trade-off.

6. Conclusion

This study presented an explainable hybrid deep–machine learning framework for predicting the Remaining Useful Life (RUL) of turbofan engines using the NASA C-MAPSS dataset. The proposed model combined a Long Short-Term Memory (LSTM) encoder for temporal feature extraction with an Extreme Gradient Boosting (XGBoost) regressor for nonlinear RUL estimation, followed by SHAP-based explainability to interpret model decisions.

The model was trained and evaluated separately on all four C-MAPSS subsets (FD001–FD004), using a structured data pipeline that integrated the official NASA training, testing, and RUL files. The framework achieved RMSE = 14.01, MAE = 9.97, and R^2 = 0.885 on FD003 comparable to or exceeding contemporary deep-learning baselines such as CNN–LSTM and Attention-LSTM models. Performance across other remained robust, confirming the model's adaptability to multi-condition and multi-fault environments. The SHAP analysis indicates that a small subset of latent

dimensions dominates RUL prediction, with Latent_10 and Latent_1 emerging as the most consistently influential across multiple datasets, followed by Latent_2 and Latent_6. Dataset-specific dominant patterns were also observed, such as Latent_3 in FD002 and Latent_11 in FD004.

References

Ahmed, A., "Hybrid and Deep Learning Architectures for Predictive Maintenance: Evaluating LSTM and Attention-Based LSTM-XGBoost on Turbofan Engine RUL," 2025.

Alomari, A., Hasan, M. K., and Fudzee, M. F. M., "Interpretable Feature Engineering and Machine Learning Framework for Remaining Useful Life Prediction Using the C-MAPSS Dataset," *Scientific Reports*, Vol. 13, 12584, 2023.

Deng, L., Ma, X., and Zhang, J., "Deep Feature Recognition-Based Remaining Useful Life Estimation Using C-MAPSS Data," *Journal of Intelligent Manufacturing*, Vol. 35, pp. 501–512, 2024.

Dida, S., Bouhidel, R., and Amirat, M., "Attention-LSTM Network for Turbofan Engine Remaining Useful Life Prediction," 2025.

Ellefsen, A. L., Bjørlykhaug, E., Æsøy, V., Ushakov, S., and Zhang, H., "Remaining Useful Life Predictions for Turbofan Engine Degradation Using LSTM Networks," 2019.

Ensarioğlu, M., Eren, S., and Türkcan, M., "Benchmarking Labeling and Preprocessing Effects on RUL Prediction for the C-MAPSS Dataset," *Applied Sciences*, Vol. 13, No. 8, 4769, 2023.

Hou, J., Zhang, X., and Jiang, L., "Deep Learning for Prognostics and Health Management: A Review with C-MAPSS Applications," *Frontiers in Artificial Intelligence*, Vol. 3, pp. 1–14, 2020.

Jean-Pierre, M., Celestin, J. R., and Smith, A. N., "Transformer-Based Ordinal Regression for Censored Remaining Useful Life Data," 2024.

Li, X., Ding, Q., and Sun, J.-Q., "Remaining Useful Life Estimation in Mechanical Systems Using Deep Convolutional Neural Networks," *Mechanical Systems and Signal Processing*, Vol. 104, pp. 799–810, 2018.

Peng, Y., Zhao, T., and Tang, D., "Temporal–Spatial Feature Fusion for Remaining Useful Life Estimation," *Frontiers in Neurorobotics*, Vol. 15, pp. 1–12, 2021.

Soualhi, A., Medjaher, T., and Zerhouni, N., "Explainable Machine Learning for Remaining Useful Life Prediction: A Case Study on C-MAPSS Data," *Engineering Applications of Artificial Intelligence*, Vol. 132, 107670, 2024.

Vishnu, C. R., Prakash, P. S., and Krishnan, S. S., "Deep Ordinal Regression and Uncertainty Quantification for Remaining Useful Life Estimation," *arXiv preprint arXiv:1905.06358*, 2019.

Wen, C., Sun, J., and Ding, Q., "An Ensemble Residual Convolutional Neural Network for Remaining Useful Life Estimation," *AIMS Electronics and Electrical Engineering*, Vol. 3, No. 2, pp. 168–186, 2019.

Xia, W., Lin, C., and Zhang, X., "Selective Ensemble Deep Neural Networks for Remaining Useful Life Prediction," *Neurocomputing*, Vol. 570, pp. 126–139, 2024.

Yu, Z., Zhang, Y., and Gao, J., "Hybrid CNN–LSTM Network for Remaining Useful Life Prediction in Turbofan Engines," *Mechanical Systems and Signal Processing*, Vol. 213, 111240, 2025.

Biographies

Asma Sardar Asma Sardar is an undergraduate Software Engineering student at Al Yamamah University, Khobar. Her primary interests include artificial intelligence, machine learning, and intelligent systems. She has actively participated in national and global hackathons, focusing on developing AI-driven solutions for industrial and real-world challenges. She is passionate about leveraging emerging technologies to drive data-driven innovation and promote sustainable digital transformation.

Sara M. Al Ghalayini is an Undergraduate student at Industrial Engineering Department, College of Engineering Al Yamamah University. She is currently studying at King Fahad University of Petroleum and Minerals as part of the visitor student program. Her passion for industry is reflected in her work with the Industrial Engineering and Operations management Society Al Yamamah University Chapter as head of the Research and Development team..

Saba Alkhalifah is an Undergraduate student in the Industrial Engineering Department at Al Yamamah University, KSA. She is the Founder and President of the IEOM Student Chapter at the university. Saba has been actively involved in organizing university-wide events and establishing partnerships with leading industrial organizations. Her leadership experience includes managing public relations, marketing initiatives, and industry collaborations that promote student engagement and professional growth. Her interests include quality control, operations management,

industrial systems optimization, innovation, and the integration of artificial intelligence and data-driven approaches within industrial engineering applications. She has received multiple recognitions for her contributions, including acknowledgment from the Dean of the College of Engineering and certificates of excellence and volunteer service..

Rana Alwabel is an Undergraduate student at Industrial Engineering Department, College of Engineering Al Yamamah University. She is passionate about industrial systems, optimization, and data-driven thinking. Rana also serves as the Head of Finance at the IEOM Student Chapter, supporting student engineering activities and initiatives within the industrial engineering community.

Conrado Vizcarra is a lecturer at the College of Engineering at Al Yamamah University in Al-Khobar, Saudi Arabia. With two decades of experience in education, he brings a wealth of knowledge and passion for teaching to his students. Mr. Vizcarra earned his Master's Degree in Information Technology from the University of the Cordilleras in 2010. He is currently pursuing a Doctorate degree in the same field at Saint Paul University in Tuguegarao, Philippines, further solidifying his expertise and commitment to academic excellence.

Osama T. Al Meanazel is an Associate Professor in the Industrial Engineering Department at Al Yamamah University, KSA. He earned his Ph.D. in Industrial & Systems Engineering from the State University of New York (SUNY) at Binghamton, USA, in 2013, following his M.S. in Engineering Management from the University of Sunderland, UK, and his B.S. in Industrial Engineering from the University of Jordan. Dr. Al Meanazel has held various academic positions, including Visiting Professor at Applied Science Private University, and has served as Director of the Center for Studies, Consultations, and Community Service at Hashemite University. His research interests focus on industrial and systems engineering, particularly in ergonomic risk management, production efficiency, and human factors engineering. Dr. Al Meanazel has authored numerous peer-reviewed journal articles and conference papers, receiving several awards for his work, including Best Track Paper and Teaching Excellence. He is an active member of multiple professional societies, including the Human Factors and Ergonomics Society and the Institute of Industrial and Systems Engineers.