Remaining Useful Life Estimation for Aircraft Engines with Risk-Aware Prediction Intervals via Conformalized Quantile Regression

Colby Don Robinson[1]

[1]*Tinker Air Force Base, Oklahoma City, Oklahoma, United States*
colby.robinson.2@us.af.mil

[1]*University of Central Oklahoma, Edmond, Oklahoma, United States*
crobinson58@uco.edu

ABSTRACT

In aerospace maintenance, remaining useful life (RUL) prediction is critical for flight safety, system availability, and long-term sustainment. While data-driven and machine learning (ML) approaches have improved RUL accuracy, most methods provide only point estimates and either omit uncertainty quantification (UQ) or rely on fixed, fleet-wide safety margins. Without reliable uncertainty estimates, even accurate point predictions offer limited value for safety-critical maintenance decisions.

This paper presents a comprehensive framework for RUL prediction that jointly addresses point estimation, uncertainty quantification, and aerospace risk preferences. The framework combines a gradient boosting regressor (GBR) for point predictions with asymmetric conformalized quantile regression (CQR) to produce prediction intervals that communicate uncertainty. The asymmetric formulation of CQR allocates miscoverage unequally between interval bounds to reduce the likelihood of overly optimistic predictions, thereby aligning interval construction with the preference to avoid late maintenance intervention.

The framework is evaluated on NASA's Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) benchmark dataset. Across all four benchmark subsets, the framework achieves test RMSE values of 13.26-16.85 with empirical coverage of 88-92% at 90% nominal coverage. These results demonstrate accurate point predictions and well-calibrated uncertainty intervals aligned with the requirements of safety-critical maintenance planning.

https://doi.org/10.36001/IJPHM.2026.v17i1.4724

1. INTRODUCTION

Prognostics and Health Management (PHM) integrates real-time monitoring, diagnostics, prognostics, and decision support to enable condition-based and predictive maintenance, thereby improving reliability and operational safety (Zio, 2022). In aerospace applications, PHM systems are particularly important due to the safety-critical nature of operations and the severe consequences of unexpected failures (Nguyen et al., 2019). Within PHM systems, remaining useful life (RUL) estimation serves as the foundation for proactive maintenance planning, directly linking condition monitoring to actionable maintenance decisions. Because system degradation is affected by natural variability, measurement limitations, and incomplete knowledge of system behavior, RUL estimates must be accompanied by rigorous uncertainty quantification (UQ) that reflects the risks involved in maintenance decision-making (Sankararaman & Goebel, 2020).

In aircraft engines, components exhibit complex degradation patterns influenced by operational conditions, environmental factors, and usage variability. This operational diversity combined with multiple distinct failure modes, makes RUL prediction particularly challenging. Explicit modeling of each degradation mechanism is rarely feasible in practice given the diversity of fleet configurations and operating environments. To meet regulatory and certification requirements, models must generalize across heterogeneous operating regimes without requiring detailed physics-based models for each component type (Fu & Avdelidis, 2023).

The same operational variability that complicates degradation modeling and drives prediction uncertainty also increases the stakes of maintenance decision-making. In particular, aerospace maintenance decisions exhibit inherently asymmetric risk profiles, where conservative maintenance actions (early removals) are generally preferred over late decisions that increase the risk of unexpected

failures (Rengasamy et al., 2020). However, many RUL approaches omit UQ entirely or rely on fixed, fleet-wide safety margins that ignore system-specific degradation states and risk preferences. Unreliable or absent UQ drives premature removals or pushes systems past their operational limits, with direct implications for safety, availability, and cost (Sankararaman, 2015).

In parallel, much of the recent RUL literature relies on complex and computationally intensive deep learning (DL) architectures, which tends to hinder reproducibility and deployment in resource-constrained PHM systems subject to regulatory and certification requirements (Luettig et al., 2024). These limitations motivate an approach that utilizes comparatively lightweight, well-understood ML models that are easier to integrate into existing PHM workflows.

Recently, conformal prediction has emerged as a model-agnostic framework for constructing prediction intervals with finite-sample coverage guarantees under an exchangeability assumption (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Conformalized Quantile Regression (CQR) combines conformal techniques with quantile regression to yield adaptive, data-dependent intervals that better reflect heteroscedastic uncertainty (Romano et al., 2019). Javanmardi & Hüllermeier (2023) demonstrated that conformal methods, including CQR, can produce valid prediction intervals for RUL on the C-MAPSS dataset, but focused on symmetric miscoverage and empirical coverage validation. In practice, however, maintenance planners face asymmetric costs for under- and over-estimating RUL, a consideration that symmetric miscoverage allocation does not address.

To address these limitations, this paper presents a risk-aware RUL prediction framework for turbofan engines on the C-MAPSS benchmark. The framework provides point estimates and complements them with prediction intervals to communicate uncertainty. Intervals are constructed using an asymmetric variant of CQR, which assigns different miscoverage rates to the lower and upper bounds to prioritize avoiding late maintenance actions. The result is a single framework that jointly addresses point estimation, uncertainty quantification, and the integration of aerospace risk preferences. By unifying these elements, the framework enables maintenance planners to make risk-informed decisions based on calibrated prediction intervals rather than fixed safety margins.

The remainder of this paper is organized as follows. Section 2 reviews related work in RUL prediction and uncertainty quantification. Section 3 introduces the C-MAPSS dataset and describes the RUL labeling strategy. Section 4 presents the proposed framework in detail, including preprocessing, the point prediction model, and asymmetric CQR. Section 5 details the experimental setup and evaluation protocol. Section 6 reports experimental results and analyzes point and interval performance, followed by concluding remarks in Section 7.

2. LITERATURE REVIEW

2.1. Remaining Useful Life Prediction

Traditionally, RUL prediction techniques are grouped into physics-based or data-driven methods. Physics-based approaches rely on the principles of physics and mechanistic models to simulate system degradation using domain expertise and sensor measurements (Cubillo et al., 2016; Li & Lee, 2005; Marahleh et al., 2006). Data-driven methods leverage statistics, machine learning, and artificial intelligence to learn degradation patterns directly from historical or real-time operations data. These two methodologies represent a tradeoff between interpretability and data dependence: physics-based approaches excel when a deep understanding of system behavior exists, while data-driven methods thrive with access to sufficient amounts of sensor data. Given the widespread adoption of sensor-based monitoring in modern PHM systems and the growing demand for scalable, adaptable solutions, the following review emphasizes data-driven approaches to RUL prediction, with particular attention to methods evaluated on the C-MAPSS dataset.

Within data-driven methods, a substantial body of work demonstrates that traditional ML can rival deep networks when paired with strong preprocessing and feature engineering. This is particularly attractive in aerospace contexts where interpretability, computational efficiency, and reproducibility are essential. For example, Alomari et al. (2023) combined variable-stride rolling windows, statistical feature extraction, principal component analysis (PCA), and feature selection before training gradient boosting and random forest regressors, reporting competitive performance across all four C-MAPSS subsets. Likewise, Chen et al. (2020) employed feature extraction, LASSO feature selection, and a Random Forest regressor to achieve performance on par with many DL algorithms.

Among traditional ML algorithms, tree-based ensembles are prominent and often directly compared with DL. Using a Light Gradient Boosting Machine (LightGBM), Li et al. (2018) reported lower RMSE than several deep models on FD001. Ensarioglu et al. (2023) evaluated GBR and Random Forest regressors alongside their proposed CNN–LSTM architecture on FD001; while the deep model achieved the lowest errors, ensemble learners remained competitive and offered superior computational efficiency. Lin et al. (2025) further found XGBoost to outperform Random Forest and LSTM baselines in both accuracy and training time. Hybrid approaches like Liu et al. (2021) also leverage tree-based ensembles: a convolutional neural network (CNN) was used for feature extraction, but the neural regression layer was replaced with a LightGBM regressor, resulting in improved

accuracy and robustness over the standalone deep model. Collectively, these studies highlight that carefully engineered features combined with ensemble regressors can match or exceed the performance of more complex deep architectures on C-MAPSS, while remaining computationally efficient and well-understood.

Having established the capabilities of traditional ML approaches, this section shifts its focus to DL methods, which learn degradation-related features directly from high-dimensional sensor streams. Early work applied recurrent architectures such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) to capture temporal dependencies in RUL trajectories (Zheng et al., 2017; Chen et al., 2019; Wu et al., 2020). In parallel, convolutional neural networks (CNNs) have been studied for their capacity to learn local degradation patterns from multivariate sequences. Li et al. (2018) showed that a one-dimensional CNN achieved competitive RUL accuracy while eliminating manual feature design. Hybrid strategies, such as the CNN–BiLSTM model of Xia et al. (2020), train on varying window lengths to leverage both short- and long-term degradation trends, yielding further reductions in RMSE.

Building on recurrent and convolutional baselines, attention mechanisms have been introduced to focus on informative time segments and integrate information across temporal scales. Peng et al. (2022) combined a stacked sparse autoencoder (SSAE) with an attention-augmented Echo State Network (ESN), reporting significantly lower RMSE than other deep baselines. Similarly, Elsherif et al. (2025) integrated a convolutional autoencoder (CAE) with an attention-based LSTM, achieving RMSE values in the low teens on FD001 and FD003. Furthermore, Chen (2024) paired position-sensitive self-attention with an LSTM decoder to improve long-sequence modeling.

Across both traditional ML and DL studies, C-MAPSS is typically used with a piecewise RUL labeling strategy that assumes an initial constant RUL plateau followed by approximately linear degradation. Performance is commonly reported using metrics such as RMSE and MAE on engine-level train–test splits. These works demonstrate that both traditional ML and deep models can achieve high predictive accuracy on C-MAPSS. However, they primarily focus on point estimates, with limited attention to quantifying uncertainty or integrating maintenance-oriented risk preferences. This motivates shifting the focus to the next sections on uncertainty quantification and conformal prediction.

2.2. Uncertainty Quantification in RUL Prediction

Uncertainty quantification (UQ) is the process of characterizing and propagating the sources of uncertainty in predictive models to assess the reliability of their outputs. These sources are commonly categorized as aleatoric uncertainty, which stems from inherent process variability and measurement noise, and epistemic uncertainty, which arises from limited knowledge, data sparsity, or model inadequacy. In RUL prediction, estimates are inherently uncertain due to variability in operating conditions, unobserved degradation mechanisms, and sensor noise. Consequently, UQ is essential for aerospace prognostic systems, where maintenance and safety decisions must account not only for the predicted RUL but also for the associated confidence. This perspective is consistent with early PHM studies, which highlighted that accurate point estimates alone are insufficient for guiding decision-making (Sankararaman & Goebel, 2013; Saxena et al., 2010).

Early work began addressing this need with probabilistic modeling and state-space methods. For example, particle filtering was adapted to yield predictive distributions for RUL rather than single point estimates (Zio & Peloni, 2011). Other studies incorporated inverse First-Order Reliability Method (FORM) with state-space degradation models to produce full probability distributions for predicted RUL (Sankararaman et al., 2014). These approaches provide a principled probabilistic treatment of degradation but typically require carefully specified state-space models and distributional assumptions and can become computationally demanding for complex systems or large fleets.

As data-driven RUL modeling matured, several families of UQ methods became common. Bayesian approaches were explored for their ability to provide posterior predictive distributions and associated credible intervals for RUL (Chen et al., 2023; Lin & Li, 2022; Ochella et al., 2024). While conceptually appealing, they typically rely on parametric assumptions about the degradation or noise process and can be computationally expensive to scale or deploy in real-time PHM environments. Interval and quantile predictors directly estimate conditional upper and lower bounds to form interpretable prediction intervals (Zhao et al., 2020), but the resulting intervals do not generally guarantee coverage and may become miscalibrated under distribution or covariate shift. Ensemble-based methods, such as random forests, deep ensembles, and Monte Carlo dropout, train multiple models or perform stochastic forward passes and use the dispersion of their outputs as a proxy for predictive uncertainty (Liao et al., 2018; Jiang et al., 2025; Faizanbasha & Rizwan, 2025). However, these methods rely on model diversity to produce meaningful uncertainty estimates and may remain overconfident under distribution shift when all members share similar inductive biases.

These families differ in assumptions, computational cost, and the form in which uncertainty is reported: Bayesian methods yield predictive distributions, interval and quantile approaches produce prediction intervals, and ensemble-based methods provide dispersion measures. Despite this growing body of research, the adoption of UQ in RUL applications remains limited. Many operational frameworks still rely on

fixed margins or heuristic thresholds rather than empirically calibrated intervals, and even in research settings, uncertainty quality is often evaluated only superficially. Moreover, most existing UQ methods for RUL lack formal coverage guarantees and provide no mechanism to encode asymmetric costs for under- and over-estimating RUL. These limitations motivate the use of conformal prediction techniques, which are discussed in the next subsection.

2.3. Conformal Prediction

Conformal prediction has recently emerged as a model-agnostic framework for uncertainty quantification in RUL. Conformal methods provide a distribution-free way to construct prediction intervals that are statistically valid in finite samples under an exchangeability assumption (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Unlike approaches that rely on parametric distributional assumptions, conformal prediction only requires that calibration and future test examples be exchangeable. The method uses a held-out calibration set to quantify the conformity between predictions and observations, constructing intervals that, under exchangeability, contain the true value with probability at least $1 - \alpha$ in finite samples. In the RUL setting, engine trajectories form time series with non-stationary degradation, so exchangeability is only approximately satisfied: calibration and test sets are composed of separate groups of engine trajectories, where each trajectory represents a unique non-stationary process. Though recent work has shown that split conformal prediction remains approximately valid under mild dependence (Barber et al., 2023), the coverage guarantees should be interpreted empirically on held-out engines rather than as exact in the presence of temporal dependence.

Conformalized Quantile Regression (CQR) extends the original conformal framework by leveraging quantile regression to produce more efficient intervals when the conditional distribution of the target is heteroscedastic (Romano et al., 2019). CQR first trains a quantile regression model to estimate lower and upper conditional quantiles and then applies a conformalization step to adjust these quantiles so that the resulting intervals achieve the desired marginal coverage. In contrast to fixed-width intervals, CQR adapts interval width based on the predicted quantiles, inheriting both the finite-sample validity of conformal prediction and the statistical efficiency of quantile regression. This adaptivity is particularly attractive for RUL, where predictive uncertainty varies across operating regimes and often changes throughout the degradation process.

Conformal prediction remains comparatively underexplored in prognostics. Javanmardi & Hüllermeier (2023) recently evaluated several conformal algorithms for RUL prediction on C-MAPSS, including split conformal prediction and CQR. Their results confirmed the empirical validity of conformal frameworks for RUL. Furthermore, the study demonstrated that Conformalized Quantile Regression (CQR) consistently produced narrower intervals on average compared to the other variants. This finding was based on an evaluation using standard coverage and average-width criteria under symmetric miscoverage allocation. In contrast, the present paper employs asymmetric CQR to align interval construction with aerospace maintenance risks, thereby using conformal prediction not only for statistical calibration but as a component of a risk-aware RUL framework.

3. DATASET & LABELING

3.1. C-MAPSS Dataset

The Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) dataset, released by NASA's Prognostics Center of Excellence, has become a standard benchmark for evaluating RUL prediction algorithms in aerospace applications (Ramasso & Saxena, 2014). The dataset is designed to simulate turbofan-engine degradation under varying conditions and fault modes. It is comprised of four subsets (FD001–FD004) that each contain two separate groups of engines for training and testing. Figure 1 displays a simplified diagram of the simulated turbofan-engine and its main components.

Each subset has a distinct configuration that differs by the number of fault modes, operating conditions, and sample sizes. The operational settings are designed to simulate different operating environments and represent engine conditions such as thrust level or altitude. The simulated fault modes correspond to the gradual loss of efficiency in the high-pressure compressor (HPC) and in the fan module. Table 1 summarizes how the four subsets are distinct from one another in size and configuration.

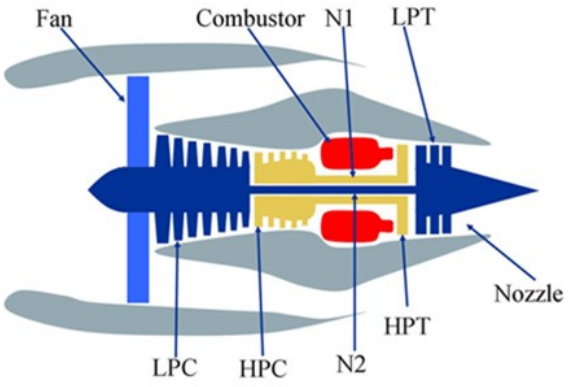

Figure 1: Simplified diagram of the turbofan-engine simulated in the C-MAPSS dataset. Reproduced from Frederick et al. (2007), NASA/TM-2007-215026.

Subset	FD001	FD002	FD003	FD004
No. of training engine trajectories	100	260	100	249
No. of test engine trajectories	100	259	100	248
Operating Conditions	1	6	1	6
Fault Modes	1	1	2	2
Training Size	20,632	53,760	24,721	61,250
Test Size	13,097	33,992	16,597	41,215

Table 1: C-MAPSS dataset summary & subset characteristics

The dataset contains 26 variables that are comprised of engine ID, engine cycle, 3 operational settings, and 21 sensor channels monitoring parameters such as temperature, pressure, and rotational speeds across engine modules. The synthetic design provides complete run-to-failure trajectories for training engines and truncated trajectories for test engines, along with ground-truth RUL values at the final observation of each test engine. It also incorporates realistic sensor noise and multiple operating conditions, enabling standardized evaluation of prognostic algorithms (Saxena et al., 2008)

3.2. RUL Labeling

The framework employs a standard piecewise RUL labeling function to represent the number of engine cycles remaining until failure. For the training set, each engine has complete run-to-failure data; for any cycle the label is determined as the number of cycles left until its last recorded observation. For the test set, the dataset provides the true RUL at each engine's final observation. To emphasize the degradation region and limit the influence of early-life cycles, RUL labels are capped at a maximum of 125 cycles.

The labeling function is defined as:

$$RUL_t = \min(RUL_{max},\, F - t)\,,\; RUL_{max} = 125 \qquad (1)$$

Where t denotes the current engine cycle and F is the engine cycle at failure.

4. PROPOSED FRAMEWORK & METHODOLOGY

This study proposes a unified framework for RUL prediction that jointly delivers point estimates, calibrated uncertainty intervals, and risk-aligned interval construction for aerospace PHM systems. The framework is composed of a systematic pipeline that transforms raw sensor streams through preprocessing, feature extraction, and dimensionality reduction stages before branching into two modeling processes: a GBR for point predictions, and a dedicated uncertainty quantification model using asymmetric CQR to produce prediction intervals. An overview of the proposed framework is shown in Figure 2.

4.1. Data Preprocessing

This section details the data preparation steps leading up to feature engineering and extraction. Before normalization, sensor channels displaying little or near-constant variance were removed to focus on sensors that capture meaningful information about the degradation process. Near-constant sensors were identified based on their variance over the training trajectories: channels whose variance fell below a small threshold were dropped, and the same set of sensors was removed from the calibration and test data. This avoids including essentially static measurements while keeping the preprocessing consistent across all splits.

4.1.1. Normalization

Normalization is tailored to the operating-condition structure of each subset. For FD001 and FD003 (single-condition subsets), each engine's sensor channels are normalized using statistics computed from its first ten cycles, which are assumed to reflect a healthy operating baseline. For FD002 and FD004 (multi-condition subsets), the data is normalized by operating regime. Operating regimes are identified by applying k-means clustering to the three operational setting features on the training engines; the clustering is fit once to obtain centroids representing the six operating regimes. Each engine observation is then assigned to its nearest centroid by Euclidean distance, and calibration and test time steps are assigned using these frozen centroids. Within each regime, regime-specific training statistics are computed for every sensor, and regime-conditional standardized values are formed. This procedure places sensor channels on a comparable scale across distinct operating conditions while preserving degradation dynamics.

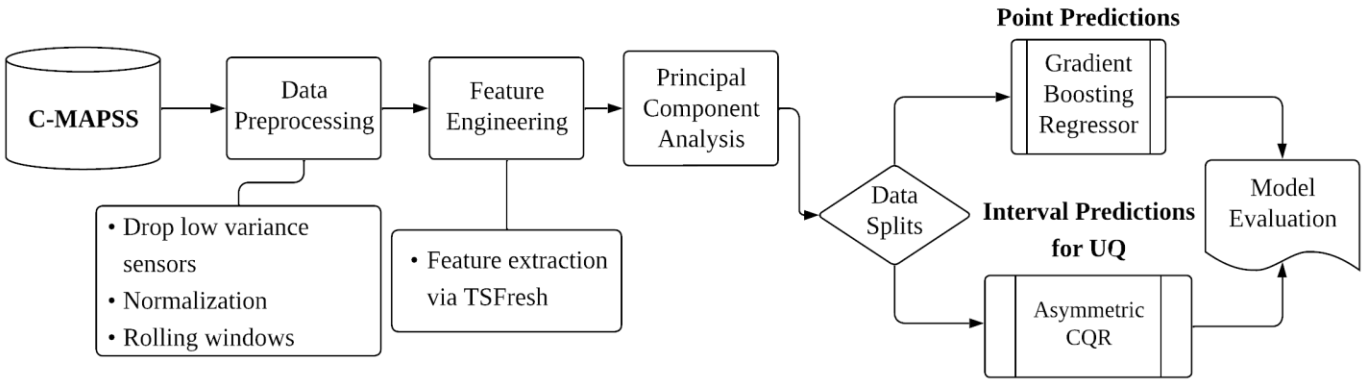

Figure 2: Flowchart of the proposed RUL prediction framework

For a given sensor value x, the standardized value is computed as

$$z = \frac{x - \mu}{\sigma + \varepsilon} \quad (2)$$

where μ and σ represent the mean and standard deviation estimated from the appropriate training reference, and ε is a small constant added to the denominator to avoid division by very small standard deviations.

4.1.2. Rolling Windows

To encode short-term temporal dependencies for RUL prediction, each engine trajectory is segmented into overlapping time windows of 30 cycles that advance forward in time. The stride between successive windows is variable, randomly selected per engine between 5 and 20 cycles, which introduces sampling diversity across engine trajectories while maintaining reproducibility through a fixed random seed. The same variable-stride windowing configuration is applied across all four subsets, and each window is treated as an independent training example. The label associated with each window is the RUL at the window's final cycle, ensuring that targets align with the decision point. Windows derived from engines in the training split are used to fit the models, while windows from calibration and test engines are reserved for conformal calibration and final evaluation, respectively.

4.2. Feature Engineering & Extraction

After normalization and windowing, features are extracted from the sensor data using TSFresh (Time Series FeatuRe Extraction on basis of Scalable Hypothesis Tests), an open-source library that automatically computes a predefined set of statistical descriptors from time-series data (Christ et al., 2018). For each 30-cycle window, TSFresh produces a fixed-length feature vector summarizing level and dispersion, local temporal structure, and spectral and entropy characteristics. This automated extraction reduces reliance on manual feature 121design. The specific feature families used in this study are listed in Table 2.

Extracted Time-Series Features

Mean	First location of maximum & minimum
Standard Deviation	Last location of maximum & minimum
Root Mean Square	Time reversal asymmetric statistic
Autocorrelation	Partial Autocorrelation
Maximum & Minimum	Third order auto-cumulant
Mean Change	Cross entropy
Linear Trend attributes *(intercept, slope, and std. error)*	
Augmented Dickey Fuller test *(test-statistic, p-value)*	
Lempel Ziv complexity *(bins: 5)*	
Permutation entropy *(dimension: 3, tau: 1)*	
Fast Fourier Transform Coefficient *(coefficients 0 to 10, attribute: absolute value)*	
Fast Fourier Transform Aggregate *(centroid, variance, skewness, and kurtosis)*	

Table 2: Time-series features extracted from C-MAPSS sensor data

4.3. Dimensionality Reduction

Following feature extraction, Principal Component Analysis (PCA) is applied to the TSFresh feature matrix to reduce dimensionality. PCA is fit only on the training features: the training matrix is centered, principal directions are learned, and the number of components is chosen so that the

cumulative explained variance reaches 50%, reflecting a deliberate tradeoff between retaining predictive signal and limiting model complexity. The resulting linear transformation is then frozen and applied unchanged to the calibration and test features, projecting all splits into the same PCA space without recomputing statistics on non-training data.

4.4. Data Partitioning

All data splits are performed at the engine level to prevent leakage between training, calibration, and test sets. For the point-prediction model, the original C-MAPSS training engines are used to fit the model, and the corresponding test engines are reserved for final evaluation. For the CQR module, the same engine-level test set is used, while the training engines are further partitioned into training and calibration subsets: for each subset, 30% of the training engines are randomly designated as calibration engines and used only to compute conformal adjustments.

4.5. Feature Selection

After PCA, Random Forest–based feature selection is applied to the resulting principal components to retain the most informative predictive signals. A Random Forest regressor is fit on the training PCA features, and component importance is estimated using mean decrease in impurity (MDI). The median of the component importance values is then used as a selection cutoff: only components with importance above the median are retained. This removes the need to tune an additional hyperparameter for the number of selected components, yielding a compact feature set for downstream models.

4.6. Point Prediction Model

The point-prediction model estimates remaining useful life as a single point value from the selected PCA components. A gradient boosting regressor is used as the point predictor because it can model nonlinear interactions among engineered features while providing regularization and efficient training. Gradient boosting constructs an ensemble of shallow regression trees in a stagewise manner, where each tree is trained to correct the residual errors of the current ensemble by following the negative gradient of the loss function. The final prediction is obtained by summing the contributions of all trees in the ensemble, scaled by a learning rate that controls the step size. This subsection establishes the point-prediction model; interval predictions are introduced separately in the following subsections via asymmetric CQR, which does not modify the point predictor or preprocessing stages.

4.7. Conformalized Quantile Regression

To complement the point predictions and provide calibrated uncertainty information, the framework employs Conformalized Quantile Regression (CQR) to produce prediction intervals (Romano et al., 2019). Two GBRs using pinball (quantile) loss functions are trained to approximate the conditional lower and upper quantiles of RUL at levels $\tau_{\text{low}} = \alpha/2$ and $\tau_{\text{high}} = 1 - \alpha/2$, where α is the target miscoverage rate. These models provide preliminary, heteroscedastic prediction intervals whose width adapts to the estimated local uncertainty.

The conformal step uses a held-out calibration set, disjoint from the training data, to correct these preliminary intervals and guarantee marginal coverage under an exchangeability assumption. For each calibration sample (x_i, y_i), nonconformity scores are computed as

$$s_i = \max\{\hat{q}_{\tau_{low}}(x_i) - y_i,\ y_i - \hat{q}_{\tau_{high}}(x_i)\}, \qquad (3)$$

Where $\hat{q}_{\tau_{low}}$ and $\hat{q}_{\tau_{high}}$ denote the predicted lower and upper quantiles, respectively. Let $\hat{Q}_{1-\alpha}$ be the empirical quantile at level $(1 - \alpha)$ of the calibration scores $\{s_i\}_{i=1}^{n_{cal}}$ with the standard finite-sample correction, obtained as

$$\hat{Q}_{1-\alpha} = s_{(\lceil (n_{cal}+1)(1-\alpha) \rceil)}, \qquad (4)$$

where $s_{(k)}$ denotes the k-th order statistic of $\{s_i\}_{i=1}^{n_{cal}}$. The final symmetric CQR interval for a new feature vector x is then

$$\left[\hat{q}_{\tau_{low}}(x) - \hat{Q}_{1-\alpha}\,, \hat{q}_{\tau_{high}}(x) + \hat{Q}_{1-\alpha}\right]. \qquad (5)$$

Under exchangeability between the calibration and test samples, these intervals achieve marginal coverage close to the nominal level $(1 - \alpha)$ while remaining adaptive to heteroscedasticity in the RUL predictions.

4.7.1. Asymmetric Miscoverage Allocation

Standard CQR allocates miscoverage symmetrically between the lower and upper tails. In RUL prediction, overestimating remaining life is typically more costly than underestimating it. To reflect this operational preference, an asymmetric allocation of miscoverage across the two tails is introduced.

Given a total miscoverage level α and an asymmetry ratio $r > 1$, miscoverage is split asymmetrically between upper and lower tails. This procedure allocates $\alpha_{low} = \alpha/(1 + r)$ to lower-tail miscoverage and $\alpha_{high} = \alpha r/(1 + r)$ to upper-tail miscoverage ($\alpha_{low} + \alpha_{high} = \alpha$).

The nominal quantile levels are defined as

$$\tau_{\text{low}} = \frac{\alpha}{1+r}, \tau_{\text{high}} = 1 - \alpha \cdot \frac{r}{1+r}. \qquad (6)$$

The calibration procedure is adapted by computing separate calibration residuals for the lower and upper tails. For each calibration sample (x_i, y_i),

$$s_{\text{low},i} = \{\hat{q}_{\tau_{low}}(x_i) - y_i\}, \quad (7)$$

$$s_{\text{high},i} = \{y_i - \hat{q}_{\tau_{high}}(x_i)\}, \quad (8)$$

where $\hat{q}_{\tau_{low}}$ and $\hat{q}_{\tau_{high}}$ are the preliminary lower and upper quantile predictors at levels τ_{low} and τ_{high}.

Let $\hat{Q}_{low}$ and $\hat{Q}_{high}$ denote the empirical quantiles of $\{s_{\text{low},i}\}$ and $\{s_{\text{high},i}\}$, respectively, chosen so that lower- and upper-tail miscoverage on the calibration set are controlled at $\alpha_{low}$and α_{high}. The final asymmetric prediction interval for a new feature vector x is

$$\left[\hat{q}_{\tau_{low}}(x) - \hat{Q}_{low}, \hat{q}_{\tau_{high}}(x) + \hat{Q}_{high}\right]. \quad (9)$$

This asymmetric construction discourages overly optimistic RUL estimates by reducing lower-tail miscoverage, while permitting more frequent upper-tail violations to reflect a conservative preference for underprediction (earlier intervention) in safety-critical maintenance decisions. The degree of this asymmetry is controlled by the ratio r, which reflects a tradeoff between risk reduction and interval efficiency: increasing r improves coverage for the lower tail, but at the cost of wider prediction intervals.

5. EXPERIMENTAL SETUP

All experiments are conducted in Python, with the scikit-learn library used for model training. The framework is trained and evaluated on the official training and test partitions for each C-MAPSS subset as provided by NASA. For all experiments, $\alpha = 0.1$, targeting nominal marginal coverage of $1 - \alpha = 0.9$. Additionally, the random seed is fixed for all stochastic components to ensure reproducibility, and the implementation code is publicly available on GitHub[2].

5.1. Hyperparameter Tuning

Hyperparameters for the point predictor are selected via randomized search with 5-fold cross-validation, scored by mean absolute error. The configuration with the lowest average validation error is selected, after which the model is refit on the full training set and held fixed for downstream evaluation. The quantile regressors used in CQR employ a fixed set of hyperparameters chosen to balance model capacity with stable quantile estimation.

The asymmetry ratio r is selected for each subset via cross-validation from the candidate set $r \in \{2, 3, 4\}$. The selected r is the candidate whose reduction in lower tail violations is largest relative to the corresponding increase in interval width when compared against the equal allocation case ($r = 1$).

[2]https://github.com/ColbyRobinson/RUL-Prediction-Framework-with-Risk-Aware-Prediction-Intervals-via-Conformalized-Quantile-Regression

This selection is performed entirely on training data, leaving the calibration set reserved exclusively for conformal adjustment. Applying this procedure yields $r = 2$ for FD001 and FD003 and $r = 3$ for FD002 and FD004. Table 3 summarizes the search space for the point predictor and the fixed values for the quantile regressors.

Hyperparameters	Point Predictor	Quantile Regressors
Loss function	squared error	quantile
No. of trees	{100, 200, 300}	200
Maximum depth	{3, 5, 7}	5
Learning rate	{0.01, 0.05, 0.1}	0.05
Subsample Fraction	{0.8, 0.9, 1.0}	1.0
Min. samples per split	{2, 5, 10}	5
Min. samples per leaf	{2, 4}	2
Max. features per split	{$\sqrt{n}$, 0.8, 1.0}	1.0

Table 3: Hyperparameter tuning configuration for each modeling component

5.2. Evaluation Protocol & Metrics

For point predictions, the framework is evaluated on the final window of each engine using Root Mean Squared Error (RMSE) and the scoring function from the PHM08 challenge (Saxena et al., 2008). RMSE is the standard metric used in C-MAPSS literature for point estimates and is computed as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}, \quad (10)$$

where n is the number of evaluated predictions and d_i is the residual for prediction i. The PHM08 scoring function applies asymmetric exponential penalties, penalizing late predictions more heavily than early ones. For each prediction, the residual is defined as

$$\text{s} = \sum_{i=1}^{n} s_i,$$

$$s_i = \begin{cases} \exp\left(\frac{-d_i}{13}\right) - 1, d_i < 0 \\ \exp\left(\frac{d_i}{10}\right) - 1, d_i \geq 0 \end{cases}, \quad (11)$$

where s is the computed score, n is the number of evaluated predictions, and d is the residual. Positive residuals are penalized more strongly than negative residuals, consistent with the original PHM08 challenge definition.

To evaluate the quality of interval predictions, three metrics are used. Prediction Interval Coverage Probability (PICP) measures the fraction of true RUL values contained within the prediction intervals. PICP is defined as

$$PICP = \frac{1}{n}\sum_{i=1}^{n} 1\{y_i \in [\hat{y}_i^L, \hat{y}_i^U]\}, \quad (12)$$

where n denotes the number of evaluated samples, y_i is the true RUL value for sample i, and $\hat{y}_i^L$ and $\hat{y}_i^U$ are the corresponding lower and upper bounds of the prediction interval. Mean Prediction Interval Width (MPIW) quantifies the average width of the predicted intervals over the evaluation set. MPIW is defined as

$$MPIW = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^U - \hat{y}_i^L), \quad (13)$$

where n, $\hat{y}_i^L$, and $\hat{y}_i^U$ are as previously defined in Eq. (12). Prediction Interval Normalized Average Width (PINAW) represents the average prediction interval width normalized by the scale of the target variable, allowing interval sharpness to be interpreted relative to the outcome range. PINAW is defined as

$$PINAW = \frac{1}{nR}\sum_{i=1}^{n}(\hat{y}_i^U - \hat{y}_i^L), \quad R = y_{max} - y_{min}, \quad (14)$$

where all terms are as previously defined in Eq. (12), and $R = y_{max} - y_{min}$ represents the range of true RUL values used for normalization. In addition, empirical conditional coverage is computed across binned true RUL ranges using the final-window RUL of each test engine to assess how reliably the intervals perform at different stages of the degradation process.

6. RESULTS

This section presents the experimental results for the proposed framework on all four C-MAPSS subsets. Figure 3 shows the final-window point predictions and CQR prediction intervals compared to true RUL for test engines in each subset. Engines are sorted by descending true RUL on the x-axis; the y-axis represents remaining life in cycles.

6.1. Point Prediction Performance

Table 4 presents the final evaluation metrics for the point predictor on each subset. For the single operating condition subsets (FD001/FD003), the model achieves RMSE values of 13.26 and 13.73, with PHM08 scores of 301.1 and 343.15. For the multiple operating condition subsets (FD002/FD004), the model yields RMSE values of 15.37 and 16.85, with PHM08 scores of 1151.6 and 1920.5. Across all subsets, the coefficient of determination (R^2) lies between 0.84 and 0.89, indicating that the model explains most of the variance in final-window RUL. The modest increase in RMSE from single-condition to multi-condition subsets suggests that the model maintains relatively stable accuracy as operating conditions and fault types become more complex. The larger PHM08 scores on FD002 and FD004, together with their higher RMSE and slightly lower R^2, suggests that these multi-condition subsets remain more challenging due to their greater operating variability and multiple fault modes.

Subset	RMSE	Score	R^2
FD001	13.26	301.1	0.89
FD002	15.37	1151.6	0.87
FD003	13.73	343.15	0.88
FD004	16.85	1920.5	0.84

Table 4: Point-prediction performance results for each subset of C-MAPSS

6.2. Interval Prediction Performance

Table 5 reports the interval prediction metrics obtained with asymmetric CQR on each subset. For the single operating condition subsets, empirical coverages are 0.92 and 0.88 with corresponding MPIW values of 47.6 and 42.2. For the multiple operating condition subsets, empirical coverages are 0.91 and 0.92 with MPIW values of 50.7 and 58.5.

To further evaluate the performance of the prediction intervals, Table 6 reports the empirical conditional coverage across final-window RUL ranges. Coverage generally remains near the nominal target, with most bins falling between approximately 0.87 and 1.00. The most pronounced under-coverage occurs for FD002 in the (20, 40] RUL range, with additional mild under-coverage appearing in two bins early in the degradation phase for FD001 and FD003.

Subset	PICP	MPIW	PINAW
FD001	0.92	47.6	0.403
FD002	0.91	50.7	0.426
FD003	0.88	42.2	0.354
FD004	0.92	58.5	0.492

Table 5: Interval prediction results for FD001-FD004

Subset	Empirical Conditional Coverage (RUL Ranges)				
	(0, 20)	(20, 40)	(40, 60)	(60, 80)	(80, 100)
FD001	0.92	0.87	0.90	0.83	1.00
FD002	0.87	0.77	0.89	1.00	0.94
FD003	1.00	0.89	0.90	0.91	0.79
FD004	0.97	0.90	0.90	0.95	0.93

Table 6: Empirical conditional coverage of prediction intervals across binned ranges of final-window RUL

Figure 4 illustrates per-engine degradation curves for four representative test engines, showing the true RUL, point predictions, and the corresponding asymmetric CQR intervals. In these examples, the intervals are narrow where the model is more confident and widen in regions of higher uncertainty. As degradation patterns become more pronounced, the intervals often tighten near end of life, reflecting the reduced uncertainty about the remaining cycles.

6.3. Ablation Study

To isolate the contribution of asymmetric miscoverage allocation, the proposed asymmetric CQR formulation is compared with a standard symmetric CQR baseline. As established in Section 4.7.1, the asymmetric variant redistributes allowable miscoverage between tails because interval violations have fundamentally different consequences in aerospace maintenance: lower-tail violations imply that an engine may fail earlier than anticipated and therefore pose a direct safety risk, whereas upper-tail violations primarily result in conservative early maintenance and added cost. To ensure that any observed differences in interval performance are attributable only to this quantile allocation strategy, both variants target 90% nominal coverage and all other pipeline components are held constant.

Table 7 compares symmetric and asymmetric CQR across all four C-MAPSS subsets in terms of empirical coverage, interval width, and tail-specific violation frequency. Across all four subsets, asymmetric quantile allocation reduces lower-tail violations by approximately 37–68%, yielding an overall reduction of 51%. PICP remains within two percentage points of the symmetric baseline, indicating that overall coverage is preserved even as the remaining miscoverage is redistributed. This shift is reflected in the tail-specific violation rates, which show that interval error is reallocated from the more consequential lower tail to the upper tail. This distinction is important because two methods with similar empirical coverage can still produce materially different operational risk profiles depending on where the remaining violations occur.

While the approach reduces lower-tail violations, it also produces moderately wider prediction intervals that reflect the conservative shift in the lower bound. The corresponding increase in interval width, while potentially resulting in earlier maintenance actions, is proportionally small relative to the reduction in lower tail violations. The clearest example is FD004, where asymmetric CQR reduced lower-tail violations from 16 to 5 while increasing MPIW by only 4%, demonstrating that substantial safety gains can be achieved with minimal loss of interval precision on complex, multi-condition datasets.

Subset	CQR Method	PICP	MPIW	PINAW	Lower-tail Violation Rate	Upper-tail Violation Rate
FD001	Symmetric	0.93	46.7	0.395	5% (5)	2% (2)
	Asymmetric	0.92	47.6	0.403	3% (3)	5% (5)
FD002	Symmetric	0.90	47.5	0.401	6.2% (16)	3.1% (8)
	Asymmetric	0.91	50.7	0.426	3.5% (9)	5.4% (14)
FD003	Symmetric	0.87	38.2	0.321	8% (8)	5% (5)
	Asymmetric	0.88	42.2	0.354	5% (5)	7% (7)
FD004	Symmetric	0.91	56.2	0.472	6.5% (16)	2.8% (7)
	Asymmetric	0.92	58.5	0.492	2% (5)	5.6% (14)

Table 7: Ablation study results for symmetric versus asymmetric CQR. Absolute lower- and upper-tail violation counts are given in parentheses alongside each rate.

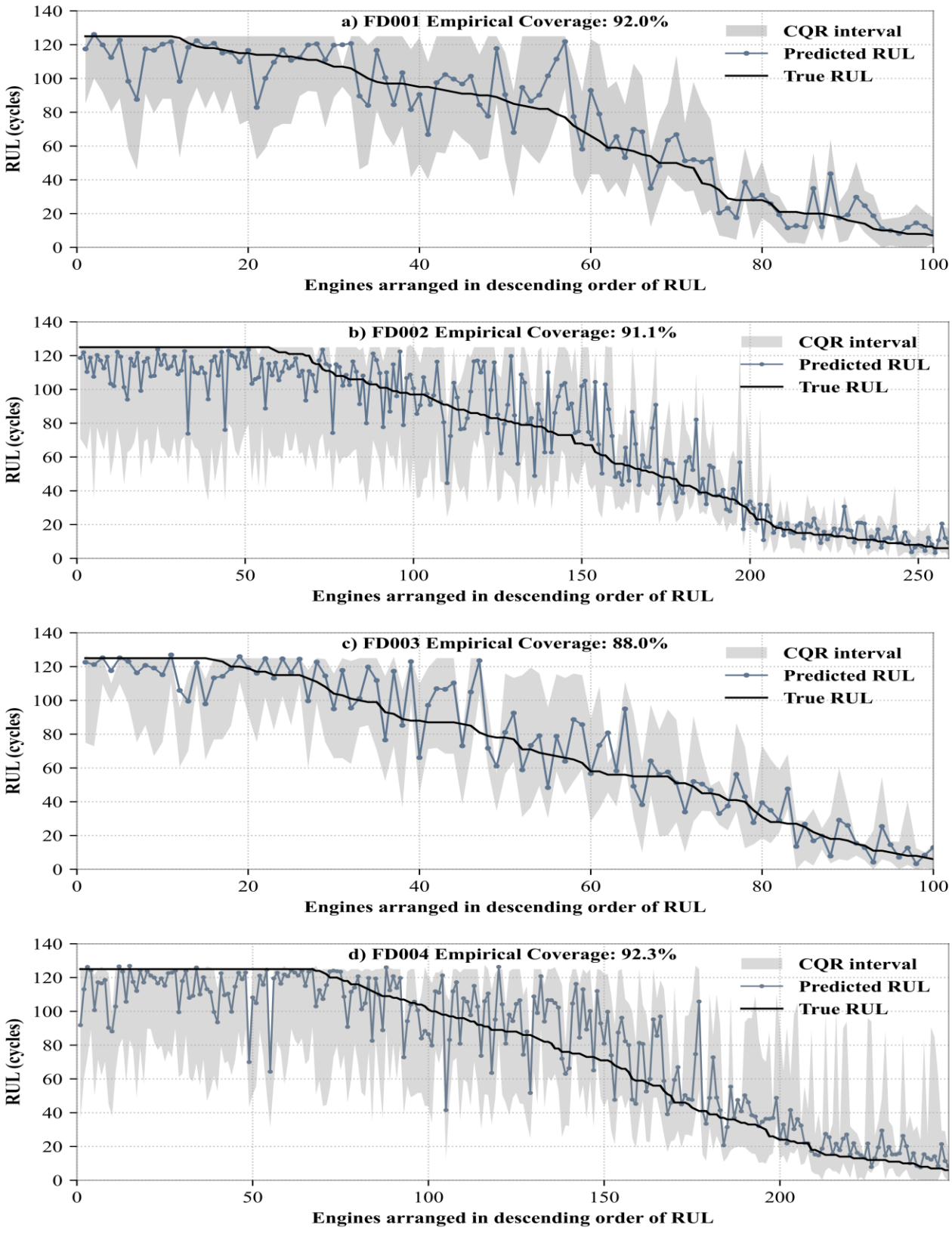

Figure 3: Final-window RUL predictions and asymmetric CQR intervals for FD001–FD004

6.4. Discussion

The experimental results demonstrate that the proposed GBR-CQR framework achieved accurate RUL estimates while providing uncertainty intervals that consistently attained near-nominal empirical coverage. Furthermore, the asymmetric CQR module was specifically designed to produce uncertainty intervals that better reflect the risk-averse nature of aerospace maintenance planning. This approach successfully reduced both the frequency and average magnitude of over-predictions compared to standard CQR.

The framework's performance is underpinned by effective preprocessing and feature engineering stages that transform raw sensor data into structured features that capture degradation progression. Trained on this engineered feature set, the GBR achieved RMSE scores between 13.26 and 16.85 across the four C-MAPSS subsets, with modest error increases as dataset complexity grew from single to multiple operating conditions and fault modes. This pattern suggests relatively stable cross-subset performance without the architectural complexity and tuning demands typically associated with deep models.

To further validate the proposed framework, its point-prediction performance is benchmarked against several DL architectures from related prognostics studies, as shown in Table 8. The comparison includes only studies reporting RMSE on all four subsets, ensuring a consistent evaluation basis. Despite differences in model architecture and training protocols across studies, two consistent patterns emerge related to operating-condition variability and fault complexity. On single operating-condition subsets (FD001/FD003), the GBR outperforms most DL models while trailing slightly behind the strongest CNN- and TCN-based approaches. The more notable contrast emerges under increased operating-condition and fault mode complexity: several DL architectures exhibit substantially higher errors on FD002 and FD004, whereas the proposed model's errors rise only modestly. Additionally, multiple deep models incur significantly larger PHM08 scores, reflecting more frequent or severe late predictions of RUL. Overall, the proposed framework consistently ranks near the top of the compared methods, attaining the second-best overall performance in terms of mean RMSE and PHM08 score among the approaches listed in Table 8. These insights show that the chosen gradient boosting backbone provides a performance level comparable to strong deep baselines, while remaining simple enough to serve as a computationally efficient foundation.

Having established the point predictor's standing relative to DL baselines, the uncertainty quantification results are now examined. Overall, the prediction intervals closely match the nominal 90% target, with empirical PICP

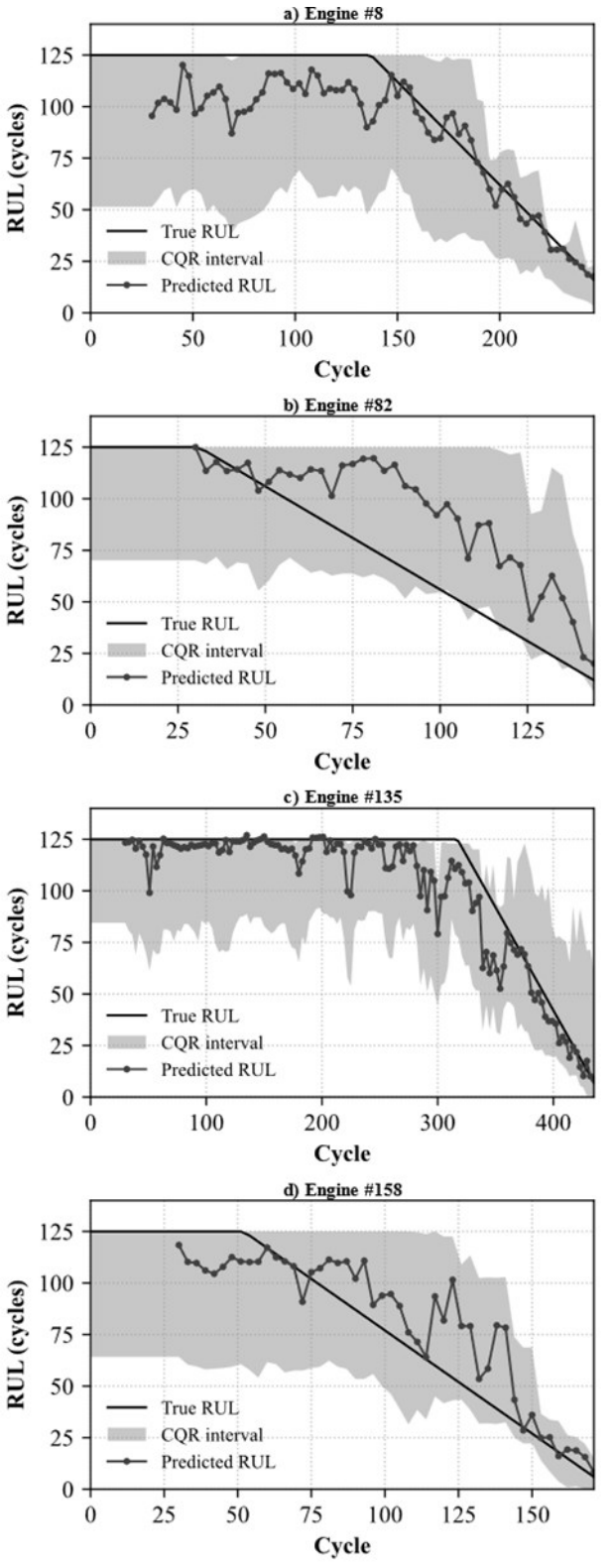

Figure 4: Individual engine degradation curves from four different engines in the FD004 test data

Method	FD001		FD002		FD003		FD004		Mean RMSE	Mean Score
	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score		
LSTM (Zheng et al., 2017)	16.14	338	24.49	4,450	16.18	852	28.17	5,550	21.25	2,797
Bi-LSTM (Wang et al., 2018)	13.65	295	23.18	4,130	13.74	317	24.86	5,430	18.86	2,543
DCNN (Li et al., 2018)	12.61	273	22.36	10,412	12.64	284	23.31	12,466	17.73	5,858
DA-TCN (Song et al., 2021)	11.78	229	16.95	1,842	11.56	257	18.23	2,317	14.63	1,161
CNN-BGRU-SA (Sun et al., 2022)	13.88	248	17.25	1,140	14.85	295	19.39	1,840	16.34	880.75
LSTM-Attention (Da Costa et al., 2023)	13.95	320	17.65	3,366	12.72	223	20.21	3,100	16.13	1,752
CNN-LSTM-Attention (Deng & Zhou, 2024)	15.97	-	14.45	-	13.90	-	16.63	-	15.24	-
Proposed GBR + CQR Framework	**13.26**	**301**	**15.37**	**1,151**	**13.73**	**343**	**16.85**	**1,920**	**14.80**	**928.81**

Table 8: Comparison of point-prediction performance between the proposed framework and multiple deep learning baselines evaluated on C-MAPSS

ranging from 0.88 to 0.92 across the four subsets—indicating that intervals are neither systematically under- nor over-confident at the marginal level. Conditional coverage by final-window RUL range is similarly consistent: most bins lie close to the nominal level, with localized under-coverage in early degradation stages offset by bins exceeding 0.90. Notably, FD004 meets or exceeds the nominal coverage level in every bin despite being the most complex subset.

Furthermore, the individual engine trajectories in Figure 4 reveal a characteristic pattern in which prediction intervals are wider during early engine life, when degradation signals are less pronounced, and narrower as engines near end of life, reflecting reduced uncertainty as failure becomes imminent. Finally, when compared to standard CQR, the asymmetric configuration reduced the frequency of lower-tail violations by 37–68% across subsets, illustrating how conformal prediction can be shaped around maintenance risk tolerances. These properties are particularly important for maintenance decisions in the aerospace industry, where practitioners must balance safety against unnecessary early removals.

7. CONCLUSION

This paper presented a unified framework for remaining useful life prediction that jointly addresses point estimation, uncertainty quantification, and aerospace risk preferences. The framework combines a GBR for point predictions with asymmetric conformalized quantile regression to produce prediction intervals aligned with the conservative risk profile of safety-critical maintenance decisions. By unifying these elements, this paper addresses the longstanding disconnect between RUL modeling outputs and the requirements of real-world maintenance decision support.

At the same time, several limitations point to opportunities for future work. The evaluation relies exclusively on the synthetic C-MAPSS dataset, and generalization to real fleet data with sensor drift, missing values, and evolving operating profiles remains untested. The conformal calibration procedure assumes a stable distribution between calibration and deployment, an assumption that may not hold under distribution shift or concept drift in operational environments. Additionally, the framework employs a single model family; alternative base learners or ensemble strategies may yield different trade-offs between point accuracy and interval quality. Future work should address these limitations by validating the framework on operational fleet data, developing adaptive conformal methods that maintain coverage under distribution shift. This integration should explore maintenance optimization models that explicitly leverage the calibrated prediction intervals for decision support.

REFERENCES

Alomari, Y., Andó, M., & Baptista, M. L. (2023). Advancing aircraft engine RUL predictions: An interpretable integrated approach of feature engineering and aggregated feature importance. *Scientific Reports*, 13(1), 13466. https://doi.org/10.1038/s41598-023-40315-1

Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. https://doi.org/10.1561/2200000101

Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability.

The Annals of Statistics, 51(2), 816–845. https://doi.org/10.1214/23-AOS2276

Chen, J., Jing, H., Chang, Y., & Liu, Q. (2019). Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process. *Reliability Engineering & System Safety*, 185, 372–382. https://doi.org/10.1016/j.ress.2019.01.006

Chen, S., He, J., Wen, P., Zhang, J., Huang, D., & Zhao, S. (2023). Remaining Useful Life Prognostics and Uncertainty Quantification for Aircraft Engines Based on Convolutional Bayesian Long Short-Term Memory Neural Network. *2023 Prognostics and Health Management Conference (PHM)*, 238–244. https://doi.org/10.1109/PHM58589.2023.00052

Chen, X. (2024). A novel transformer-based DL model enhanced by position-sensitive attention and gated hierarchical LSTM for aero-engine RUL prediction. *Scientific Reports*, 14(1), 10061. https://doi.org/10.1038/s41598-024-59095-3

Chen, X., Jin, G., Qiu, S., Lu, M., & Yu, D. (2020). Direct Remaining Useful Life Estimation Based on Random Forest Regression. *2020 Global Reliability and Prognostics and Health Management (PHM-Shanghai)*, 1–7. https://doi.org/10.1109/PHM-Shanghai49105.2020.9281004

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307, 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

Cubillo, A., Perinpanayagam, S., & Esperon-Miguez, M. (2016). A review of physics-based models in prognostics: Application to gears and bearings of rotating machinery. Advances in Mechanical Engineering, 8(8), 1–21. https://doi.org/10.1177/1687814016664660

Da Costa, P. R. O., Akçay, A., Zhang, Y., & Kaymak, U. (2023). Attention and long short-term memory network for remaining useful lifetime predictions of turbofan engine degradation. International Journal of Prognostics and Health Management, 10(4). https://doi.org/10.36001/ijphm.2019.v10i4.2623

De Giorgi, M. G., Menga, N., & Ficarella, A. (2023). Exploring Prognostic and Diagnostic Techniques for Jet Engine Health Monitoring: A Review of Degradation Mechanisms and Advanced Prediction Strategies. *Energies*, 16(6), 2711. https://doi.org/10.3390/en16062711

Deng, S., & Zhou, J. (2024). Prediction of Remaining Useful Life of Aero-engines Based on CNN-LSTM-Attention. *International Journal of Computational Intelligence Systems*, 17(1), 232. https://doi.org/10.1007/s44196-024-00639-w

Elsherif, S. M., Hafiz, B., Makhlouf, M. A., & Farouk, O. (2025). A deep learning-based prognostic approach for predicting turbofan engine degradation and remaining useful life. *Scientific Reports*, 15(1), 26251. https://doi.org/10.1038/s41598-025-09155-z

Ensarioğlu, K., İnkaya, T., & Emel, E. (2023). Remaining Useful Life Estimation of Turbofan Engines with Deep Learning Using Change-Point Detection Based Labeling and Feature Engineering. *Applied Sciences*, 13(21), 11893. https://doi.org/10.3390/app132111893

Faizanbasha, A., & Rizwan, U. (2025). Deep learning-stochastic ensemble for RUL prediction and predictive maintenance with dynamic mission abort policies. *Reliability Engineering & System Safety*, 259, 110919. https://doi.org/10.1016/j.ress.2025.110919

Frederick, D. K., DeCastro, J. A., & Litt, J. S. (2007). User's guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) (NASA/TM-2007-215026). NASA Glenn Research Center. https://ntrs.nasa.gov/citations/20070034949

Fu, S., & Avdelidis, N. P. (2023). Prognostic and Health Management of Critical Aircraft Systems and Components: An Overview. *Sensors*, 23(19), 8124. https://doi.org/10.3390/s23198124

Javanmardi, A., & Hüllermeier, E. (2023). Conformal Prediction Intervals for Remaining Useful Lifetime Estimation. *International Journal of Prognostics and Health Management*, 14(2). https://doi.org/10.36001/ijphm.2023.v14i2.3417

Jiang, L., Zhang, X., Cao, H., & Zhang, Y. (2025). A transformer-based framework with historical data fusion for RUL prediction. *Measurement Science and Technology*, 36(10), 106103. https://doi.org/10.1088/1361-6501/ae09c2

Li, F., Zhang, L., Chen, B., Gao, D., Cheng, Y., Zhang, X., Yang, Y., Gao, K., Huang, Z., & Peng, J. (2018). A Light Gradient Boosting Machine for Remaining Useful Life Estimation of Aircraft Engines. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 3562–3567. https://doi.org/10.1109/ITSC.2018.8569801

Li, C. J., & Lee, H. (2005). Gear fatigue crack prognosis using embedded model, gear dynamic model and fracture mechanics. Mechanical Systems and Signal Processing, 19(4), 836–846. https://doi.org/10.1016/j.ymssp.2004.06.007

Li, X., Ding, Q., & Sun, J.-Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 172, 1–11. https://doi.org/10.1016/j.ress.2017.11.021

Liao, Y., Zhang, L., & Liu, C. (2018). Uncertainty Prediction of Remaining Useful Life Using Long Short-Term Memory Network Based on Bootstrap Method. *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 1–8. https://doi.org/10.1109/ICPHM.2018.8448804

Lin, K.-Y., Hong, Y.-H., Li, M.-H., Shi, Y., & Matsuno, K. (2025). Predictive maintenance in industrial systems: an

XGBoost-based approach for failure time estimation and resource optimization. *Journal of Industrial and Production Engineering*. https://doi.org/10.1080/21681015.2025.2519369

Lin, Y.-H., & Li, G.-H. (2022). A Bayesian Deep Learning Framework for RUL Prediction Incorporating Uncertainty Quantification and Calibration. *IEEE Transactions on Industrial Informatics*, 18(10), 7274–7284. https://doi.org/10.1109/TII.2022.3156965

Liu, L., Wang, L., & Yu, Z. (2021). Remaining Useful Life Estimation of Aircraft Engines Based on Deep Convolution Neural Network and LightGBM Combination Model. *International Journal of Computational Intelligence Systems*, 14(1), 165. https://doi.org/10.1007/s44196-021-00020-1

Luettig, B., Akhiat, Y., & Daw, Z. (2024). ML meets aerospace: Challenges of certifying airborne AI. *Frontiers in Aerospace Engineering*, 3, 1475139. https://doi.org/10.3389/fpace.2024.1475139

Marahleh, G., Kheder, A. R. I., & Hamad, H. F. (2006). Creep-life prediction of service-exposed turbine blades. Materials Science, 42(4), 476–481. https://doi.org/10.1007/s11003-006-0103-8

Nguyen, V. D., Kefalas, M., Yang, K., Apostolidis, A., Olhofer, M., Limmer, S., & B¨ack, T. (2019). A Review: Prognostics and Health Management in Automotive and Aerospace. *International Journal of Prognostics and Health Management*, 10(2). https://doi.org/10.36001/ijphm.2019.v10i2.2730

Ochella, S., Dinmohammadi, F., & Shafiee, M. (2024). Bayesian neural networks for uncertainty quantification in remaining useful life prediction of systems with sensor monitoring. *Advances in Mechanical Engineering*, 16(7), 16878132241239802. https://doi.org/10.1177/16878132241239802

Peng, C., Chen, Y., Gui, W., Tang, Z., & Li, C. (2022). Remaining useful life prognosis of turbofan engines based on deep feature extraction and fusion. *Scientific Reports*, 12(1), 6491. https://doi.org/10.1038/s41598-022-10191-2

Ramasso, E., & Saxena, A. (2014). Review and Analysis of Algorithmic Approaches Developed for Prognostics on CMAPSS Dataset. Annual Conference of the PHM Society, 6(1). https://doi.org/10.36001/phmconf.2014.v6i1.2512

Rengasamy, D., Rothwell, B., & Figueredo, G. P. (2020). Asymmetric Loss Functions for Deep Learning Early Predictions of Remaining Useful Life in Aerospace Gas Turbine Engines. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN48605.2020.9207051

Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. In Advances in Neural Information Processing Systems (Vol. 32, pp. 3543–3553). https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html

Sankararaman, S. (2015). Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing*, 52–53, 228–247. https://doi.org/10.1016/j.ymssp.2014.05.029

Sankararaman, S., Daigle, M. J., & Goebel, K. (2014). Uncertainty Quantification in Remaining Useful Life Prediction Using First-Order Reliability Methods. *IEEE Transactions on Reliability*, 63(2), 603–619. https://doi.org/10.1109/TR.2014.2313801

Sankararaman, S., & Goebel, K. (2013). Why is the Remaining Useful Life Prediction Uncertain? *Annual Conference of the PHM Society*, 5(1). https://doi.org/10.36001/phmconf.2013.v5i1.2263

Sankararaman, S., & Goebel, K. (2020). Uncertainty in Prognostics and Systems Health Management. *International Journal of Prognostics and Health Management*, 6(4). https://doi.org/10.36001/ijphm.2015.v6i4.2319

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for evaluating performance of prognostic techniques. *2008 International Conference on Prognostics and Health Management*, 1–17. https://doi.org/10.1109/PHM.2008.4711436

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Evaluating prognostics performance for algorithms incorporating uncertainty estimates. *2010 IEEE Aerospace Conference*, 1–11. https://doi.org/10.1109/AERO.2010.5446828

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421. https://doi.org/10.5555/1390681.1390693

Song, Y., Gao, S., Li, Y., Jia, L., Li, Q., & Pang, F. (2021). Distributed Attention-Based Temporal Convolutional Network for Remaining Useful Life Prediction. *IEEE Internet of Things Journal*, 8(12), 9594–9602. https://doi.org/10.1109/JIOT.2020.3004452

Sun, J., Zheng, L., Huang, Y., & Ge, Y. (2022). Remaining Useful Life Prediction Based on CNN-BGRU-SA. *Journal of Physics: Conference Series*, 2405(1), 012007. https://doi.org/10.1088/1742-6596/2405/1/012007

Wang, J., Wen, G., Yang, S., & Liu, Y. (2018). Remaining Useful Life Estimation in Prognostics Using Deep Bidirectional LSTM Neural Network. *2018 Prognostics and System Health Management Conference (PHM-Chongqing)*, 1037–1042. https://doi.org/10.1109/PHM-Chongqing.2018.00184

Wu, J., Hu, K., Cheng, Y., Zhu, H., Shao, X., & Wang, Y. (2020). Data-driven remaining useful life prediction via multiple sensor signals and deep long short-term memory neural network. *ISA Transactions*, 97, 241–250. https://doi.org/10.1016/j.isatra.2019.07.004

Xia, T., Song, Y., Zheng, Y., Pan, E., & Xi, L. (2020). An ensemble framework based on convolutional bi-directional LSTM with multiple time windows for remaining useful life estimation. *Computers in Industry*, 115, 103182. https://doi.org/10.1016/j.compind.2019.103182

Zhao, Z., Wu, J., Wong, D., Sun, C., & Yan, R. (2020). Probabilistic Remaining Useful Life Prediction Based on Deep Convolutional Neural Network. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3717738

Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long Short-Term Memory Network for Remaining Useful Life estimation. *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 88–95. https://doi.org/10.1109/ICPHM.2017.7998311

Zio, E. (2022). Prognostics and Health Management (PHM): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, 218, 108119. https://doi.org/10.1016/j.ress.2021.108119

Zio, E., & Peloni, G. (2011). Particle filtering prognostic estimation of the remaining useful life of nonlinear components. *Reliability Engineering & System Safety*, 96(3), 403–409. https://doi.org/10.1016/j.ress.2010.08.009