iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84520

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

Trustworthy Remaining Useful Life Prediction on NASA C-MAPSS: Coupling Conformal Uncertainty with Audited SHAP Explanations

Touseef Ahmed[1], Long Wang[2]
School of Mechatronics Engineering, Anhui University of Science and Technology, Huainan, Anhui, China

Highlights

- A post-processing trustworthiness layer for RUL models that requires no retraining
- CQR restores 89.3% coverage and adapts interval width 2.8-fold across life stages
- A two-part audit confirms SHAP sensor rankings are faithful to the model
- Conformal interval width correlates with SHAP instability ($\rho = 0.733$, $n = 300$)
- Interval width acts as a single distribution-free trust signal for maintenance

Abstract: ***Remaining useful life (RUL) prediction on NASA C-MAPSS is mature, yet two requirements of safety-critical deployment remain neglected: calibrated uncertainty quantification and verification, rather than mere reporting, of model explanations. This work adds a trustworthiness layer to standard degradation models as pure post-processing, without retraining. Split conformal prediction and conformalized quantile regression (CQR) are benchmarked on gradient boosting and LSTM models trained on FD001. Fixed-width conformal intervals reach 84.7% coverage at the 90% nominal level and are miscalibrated across the lifetime; CQR restores 89.3% coverage, narrows mean width from 60.3 to 56.1 cycles, and adapts width 2.8-fold. A two-part audit confirms SHAP attributions are faithful: permuting the three top-ranked sensors raises validation RMSE by 8.22 cycles versus 0.65 for the three lowest-ranked, and SHAP agrees with permutation importance ($\rho = 0.868$). Centrally, per-instance attribution instability correlates strongly with interval width ($\rho = 0.733$, $n = 300$), so conformal interval width serves as a single distribution-free trust signal flagging unreliable predictions and explanations alike.***

Keywords: remaining useful life; conformal prediction; SHAP; uncertainty quantification; explainability; predictive maintenance.

I. INTRODUCTION

Machine elements wear out gradually in service: inside an aircraft gas turbine engine, degradation translates into slowly accumulating losses of compressor flow capacity and turbine efficiency, diminishing stall margins and cutting into operating temperature margins until safe operation is no longer assured [1]. Predicting remaining useful life (RUL) of machinery based on sensor measurements is therefore one of modern mechanical engineering's central challenges: an accurate forecast enables maintenance scheduling to avoid failure while eliminating unnecessary interventions and has immediate benefits in safety and fleet readiness. The NASA commercial C-MAPSS turbofan degradation dataset [1,2], generated by simulating a thermodynamic cycle of a high-bypass turbofan in the 90,000 lb thrust class, has emerged over the last decade as the standard problem on which data-driven machine-learning prognostics algorithms are developed and evaluated. Recurrent neural networks [3,4], ensemble learners [5], and more recently deep convolutional architectures [6] have pushed predictive RMSE on this dataset to the point where further improvements are unlikely to justify the effort required. The bottlenecks now lie elsewhere.

Specifically, two shortcomings motivate the present contribution. The first is uncertainty: point predictions of RUL, however statistically accurate on average, are of limited use for maintenance decisions with asymmetric failure cost; instead the decision maker needs a prediction interval with a well-calibrated error rate. The second is explanation: increasingly, post-hoc explanations are attached to RUL predictions in the form of feature attribution; more specifically, sensors are ranked according to their SHAP importance [7,8], yet this ranking is reported in many studies without any verification that it matches the model's understanding of the data.

Both aspects of model trustworthiness, namely the confidence in a prediction and the clarity with which it can be explained, are currently judged based on unaudited heuristics: confidence intervals on RUL are often reported as ± RMSE, with no stated error rate, while feature attributions are accepted at face value. The nascent field of explainable AI has spent the last several years showing why this should not be so: attribution methods for neural networks have been shown to be sensitive to small input perturbations [9], fail basic sanity checks [10], and be susceptible to active manipulation [11]. The same lack of due diligence applies to uncertainty estimates, which are rarely tested for calibration.

This paper makes the case that the two should be studied together, and contributes initial answers to the question of how uncertainty in the model's prediction relates to uncertainties in its explanation. Research on the C-MAPSS dataset has so far addressed uncertainty quantification and explanation fidelity at arm's length: uncertainty estimation attaches conformal or Bayesian intervals to existing RUL predictors, while explainability has been concerned with improving and auditing post-hoc explanations such as SHAP or attention weights. Little work has examined the relation between the two. Do models that are uncertain about their predictions also provide unclear explanations? If so, the practical upside is large: an engineer monitoring engine health would only need to heed a single warning signal, say the width of the prediction interval, to know when not to trust both the prediction and its sensor-level explanation. Absent such a warning, it would be reasonable to assume that the model sees the data as clearly as it does on average.

We formalize the question and explore it empirically on C-MAPSS FD001. Three research questions guide the empirical analysis: RQ1 (Calibration): does conformal prediction produce RUL prediction intervals whose validity holds uniformly over the engine life? RQ2 (Faithfulness): does SHAP faithfully describe the model's sensor-level dependence, as measured by both input perturbation and agreement with permutation importance? RQ3 (Agreement): can we detect when conformal prediction is signaling forecast uncertainty due to confused or conflicted model explanations? This paper makes four contributions to answer these questions. First, we establish a completely leakage-controlled and reproducible FD001 benchmark including both engine-level train/test splits and baseline gradient boosting and LSTM models that are competitive with published results for their respective model classes. Second, we evaluate split conformal prediction and conformalized quantile regression (CQR) both globally (does average coverage match nominal?) and locally (does coverage vary by life stage?). Third, we perform the first quantitative faithfulness audit of SHAP attributions on this dataset. Finally, we introduce a novel per-instance measure of attribution instability and demonstrate statistically significant correlation with conformal interval width, which is, to the best of the authors' knowledge, the first quantitative study of uncertainty–explanation agreement on this dataset.

A. Related Work

The C-MAPSS dataset was constructed by inducing exponential efficiency and flow losses to modules of NASA's Commercial Modular Aero-Propulsion System Simulation until the engine reached zero on a health index [1,2]. Since its creation for a data competition in 2008, in which recurrent neural networks already featured prominently [1,3], the dataset has given rise to an extensive literature: long short-term memory (LSTM) networks [4], deep convolutional networks [6], as well as many hybrid and attention-based iterations of these models. Two conventions from this literature are used here due to strong empirical support: the piecewise-linear RUL target which clips the training label at a constant value during early, healthy operation of the engine, and performance evaluation by both root-mean-square error (RMSE) and the asymmetric scoring function introduced with the dataset [1] which more harshly penalizes late predictions than early ones.

SHAP [7] computes a Shapley value for each input feature representing its contribution to a particular prediction, and the polynomial-time TreeExplainer [8] algorithm has made exact computation of SHAP values for tree ensembles standard practice. In machine-condition monitoring, rankings of sensors according to their SHAP importance scores on C-MAPSS have started appearing regularly; these typically agree in highlighting high-pressure-compressor-related channels as important. However, other recent work in the interpretability community has shown that attributions should not be trusted blindly: Alvarez-Melis and Jaakkola [9] found that several popular attribution methods can vary wildly under input perturbations imperceptible to humans, Adebayo et al. [10] found that some saliency methods are unaffected by randomized model weights and data, and Slack et al. [11] produced adversarial models whose SHAP attributions hide their decision logic. These results inspire the faithfulness audit described in Section 2.9: rather than taking an attribution ranking as an end-product, we instead treat it as a hypothesis to be tested. Closest in spirit to our RQ2, Baptista et al. [12] correlated SHAP values with established prognostic metrics of monotonicity, trendability, and prognosability on run-to-failure engine data, although without auditing attribution faithfulness or connecting attributions to calibrated uncertainty.

Conformal prediction [13,14] transforms any point predictor into a set-valued predictor that enjoys a finite-sample, distribution-free coverage guarantee under only exchangeability between calibration and test data. The split (inductive) version of conformal prediction uses only a single held-out calibration set and requires no retraining of the model [14,15]. The key drawback of split conformal prediction, a constant interval width across the entire input space, is overcome by conformalized quantile regression [16], which conformalizes a pair of estimated conditional quantiles. Doing so both adapts the width of intervals locally (inherited from the conditional quantile estimates) and maintains the marginal coverage guarantee. Conformal methods have only recently started appearing in the engineering reliability literature; in particular, Javanmardi and Hüllermeier [17] turned point RUL estimators on C-MAPSS into valid interval predictors using several conformal variants. An attractive feature of conformal prediction's accuracy guarantee is that it does not depend on the correctness of the underlying predictive model. This matches well with safety-critical maintenance decisions. In the present work the MAPIE library [18] was used for the tree-based models and a direct implementation for the sequence model. In the literature surveyed here, uncertainty quantification and explanation have been investigated as orthogonal add-ons to some life-prediction model. Papers that report prediction intervals do not study their explanations; papers that report SHAP rankings do not quantify uncertainty and rarely validate the ranking itself. To the authors' knowledge, no prior work on C-MAPSS has related calibrated uncertainty to attribution stability at the level of individual predictions. Defining and quantifying this relationship is our RQ3.

II. MATERIAL AND METHODS

A. The C-MAPSS FD001 subset

FD001 contains run-to-failure trajectories from 100 training engines and truncated trajectories from 100 test engines (Fig. 1). Each engine operates under one flight condition with one fault mode (high-pressure compressor degradation). The records contain engine ID number, operating cycle, three operational settings, and 21 sensor channels. There are 20,631 rows of training data and 13,096 rows of test data. Training engines ran between 128 and 362 cycles before failure. The actual RUL for each of the test engines at the final cycle is provided separately and can be used as ground truth. Failure was defined as the time when the engine's health index (defined as the minimum of several operability margins, including the stall margins and the exhaust-gas-temperature margin) reached zero due to forced deterioration by exponential growth of efficiency and flow losses in the HPC module [1].

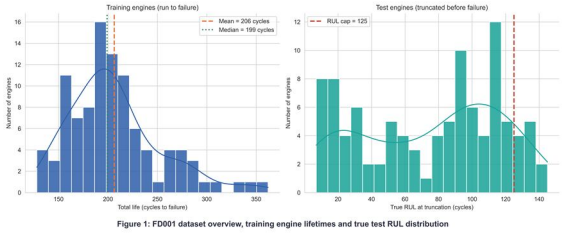


Fig. 1. FD001 overview: distribution of training-engine lifetimes (run to failure) and of true RUL for the truncated test engines.

B. Feature selection

Inspection of the per-sensor variance revealed that six channels (corresponding to sensors 1, 5, 10, 16, 18 and 19) are invariant across the whole training set and that sensor 6 has near-constant readings (standard deviation ≈ 1.4 × 10−3) (Fig. 2). This is routinely observed in the FD001 literature. These seven channels are discarded, leaving 14 informative sensors. Since FD001 is collected under only one flight condition (with the third operational setting held fixed at 100 and the first two operational settings restricted to lie in a noise band of ±0.009), the three operational settings are also discarded. Therefore all models are trained and tested on the feature vector $x \in \mathbb{R}14$.

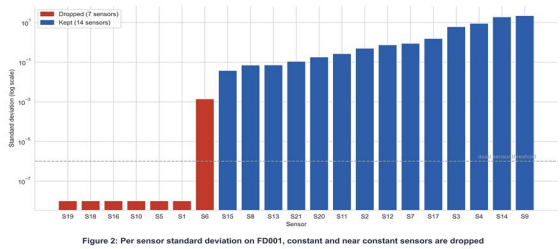


Fig. 2. Per-sensor standard deviation on FD001; the seven constant and near-constant channels are removed, retaining 14 informative sensors.

C. RUL target construction

For engine i in training with failure time Ti (maximum cycle observed for that engine), the label at cycle t is given by the usual piecewise-linear convention [1,3] (see Fig. 3),

$$RUL_i(t) = min(T_i - t, R_{max}),\ R_{max} = 125 \qquad (1)$$

This cap reflects the physical reality that degradation is inaccessible during the long healthy-life portion of the trajectory, so labels past the cap provide no useful signal. The cap is only used to shape the training target; evaluation on held-out test engines in Section 3 is reported against the raw, uncapped ground truth, which is a common convention on this benchmark.

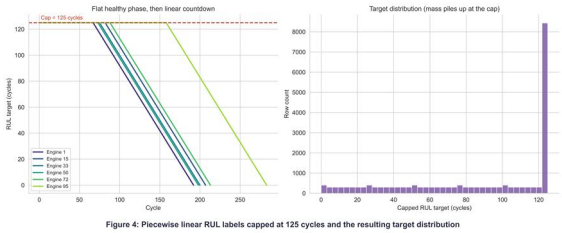


Figure 4: Piecewise linear RUL labels capped at 125 cycles and the resulting target distribution

Fig. 3. Piecewise-linear RUL target with cap of 125 cycles for representative engines, and the resulting target distribution.

D. Scaling and Data Partitioning

Features are normalized using min–max scaling fit only on the training data,

$$\tilde{x}_j = \frac{x_j - x_{j,min}}{x_{j,max} - x_{j,min}} \qquad (2)$$

where xj,min and xj,max denote the minimum and maximum of feature j over the training data, and all data splits are done at the engine level as opposed to the row level, since consecutive cycles of a given engine are highly autocorrelated and would violate the independence of partitions if split at the row level. Two splits are used: for training the reference models, the set of 100 training engines is split 80/20 into training (16,561 rows) and validation (4,070 rows) data. For running the conformal experiments, which must be done with a calibration set held out from model training, the engines are instead split 60/20/20 into training (12,270 rows), calibration (4,262 rows), and validation (4,099 rows). The held-out test set of 100 engines is used for final evaluation in both experiments.

E. Evaluation Metrics

Point accuracy is measured by the root-mean-square error,

$$RMSE = [\frac{1}{n} \Sigma_i (\hat{y}_i - y_i)^2]^{1/2} \qquad (3)$$

where ŷi and yi denote the predicted and true RUL of sample i and n is the number of evaluated samples, and by the asymmetric scoring function S introduced together with the dataset [1]. With the signed error di = ŷi − yi, the score is

$$S = \Sigma_i [exp(-d_i/a_1) - 1] \text{ for } d_i < 0;\ \Sigma_i [exp(d_i/a_2) - 1] \text{ for } d_i \geq 0 \qquad (4)$$

with a1 = 13 and a2 = 10, so that late predictions (positive di), which are dangerous in maintenance practice, are penalized exponentially more severely than early ones. Lower scores are better; this metric is denoted S throughout.

F. Reference Models

The study is anchored by two purposely distinct baseline predictors. The first is a gradient boosting regressor (GBR) [19] at its default configuration (trees = 100, max depth = 3, learning rate = 0.1). Using default hyperparameters was an explicit design choice: the goal of the GBR model is to serve as both a standard point of reference and the substrate for exact SHAP computation, while the primary contribution of this work is the trustworthiness layer applied on top of any potential baseline. The second predictor is a long short-term memory network (LSTM) [20] operating on sliding windows of width w = 30 consecutive cycles. Thus each input sample x has shape 30 × 14 and the target corresponds to the RUL value at the final cycle of each window. Note that windows are constructed within engines: the first w − 1 cycles of each engine are discarded instead of being zero-padded, as a padding value of zero corresponds to the scaled minimum across every sensor and was found to corrupt network training. This procedure results in 14,241 training windows and 3,490 validation windows under an 80/20 split.

The network consists of two LSTM layers containing 64 and 32 units respectively, followed by dropout layers of rate 0.2, and terminates in a single linear output neuron. Targets were divided by Rmax to reside in the unit interval; this was necessary for stable optimization. Training was performed with the Adam optimizer with learning rate 10−3 and batch size 64, using early stopping with patience 10 on validation loss. Early stopping restores the best weights, which were attained at epoch 19. All models were implemented in scikit-learn [21] and TensorFlow/Keras [22].

G. Split Conformal Prediction

Let f̂ be a fitted point predictor and let {(xi, yi)}, i = 1, ..., ncal, be a calibration set disjoint from the data used to fit f̂. Split conformal prediction [13-15] computes the absolute-residual nonconformity scores

$$s_i = |y_i - f(x_i)| \quad (5)$$

takes q̂ as the ⌈(ncal + 1)(1 − α)⌉ / ncal empirical quantile of these scores, and outputs the interval

$$C(x) = [f(x) - \hat{q}, f(x) + \hat{q}] \quad (6)$$

Under exchangeability of calibration and test points, this interval satisfies the finite-sample marginal guarantee

$$P(y \in C(x)) \geq 1 - \alpha \quad (7)$$

valid for any model and data distribution. Target coverage level across all inputs is 1 − α = 0.90. This guarantee is subtle: it bounds the average coverage under the input distribution rather than coverage in sub-regions of that distribution, and the width of the interval 2q̂ does not adapt to different inputs. Both of these properties are investigated empirically in Section 3.2 where we stratify coverage across bands of RUL. Recall that run-to-failure data are strongly heteroscedastic: prediction is very easy late in life when degradation signatures are strong, and genuinely difficult in the transition region corresponding to when an engine first deviates from healthy behavior.

H. Conformalized quantile regression

Conformalized quantile regression (CQR) [16] replaces the single point predictor by a pair of conditional quantile estimators q̂lo and q̂hi at levels α/2 and 1 − α/2, trained by minimizing the pinball loss

$$L_\tau(y, \hat{y}) = max\{\tau(y - \hat{y}), (\tau - 1)(y - \hat{y})\} \quad (8)$$

where τ denotes the target quantile level, which makes the raw interval [q̂lo(x), q̂hi(x)] locally adaptive: it spreads where the conditional distribution of y is wide and contracts where it is narrow. Because estimated quantiles carry no coverage guarantee of their own, CQR computes signed calibration scores

$$E_i = max\{\hat{q}_{lo}(x_i) - y_i, y_i - \hat{q}_{hi}(x_i)\} \quad (9)$$

takes Q as their ⌈(ncal + 1)(1 − α)⌉ / ncal quantile, and inflates (or deflates) the raw interval symmetrically:

$$C(x) = [\hat{q}_{lo}(x) - Q, \hat{q}_{hi}(x) + Q] \quad (10)$$

The resulting procedure maintains the marginal coverage guarantee of Eq. (7) but allows the width to change with x. Quantile estimators are gradient boosting regressors with pinball loss; the conformal correction uses the implementation of MAPIE [18]. Quantile crossing (q̂lo > q̂hi) for a small number of points, a common artifact of independently trained quantile estimators, is eliminated by post-sorting the bounds. For the LSTM, which cannot be directly used with the library interface due to its sequence input, fixed-width split conformal prediction (Eqs. 5–6) is coded explicitly for the calibration windows.

I. SHAP attributions and the Faithfulness Audit

For a model f with feature set F of size M, the SHAP value of feature j at input x is the Shapley value of the conditional-expectation game [7],

$$\varphi_j(x) = \Sigma_{S \subseteq F \setminus \{j\}} \frac{|S|!\ (M - |S| - 1)!}{M!} [f_x(S \cup \{j\}) - f_x(S)] \quad (11)$$

computed here exactly with TreeExplainer [8] on the GBR, over a random sample of 1,000 validation rows. Global importance is the mean absolute SHAP value per feature. Two independent tests then audit the resulting ranking.

Perturbation test. If the ranking is faithful, destroying the information in the top-ranked sensors must damage the model far more than destroying that in the bottom-ranked sensors. A feature subset A is corrupted by permuting each of its columns independently across the validation rows, which destroys the feature–target relationship while preserving each marginal distribution, and the damage is the resulting error increase,

$$\Delta RMSE(A) = RMSE(X_{perm(A)}) - RMSE(X) \quad (12)$$

The test compares ΔRMSE for the top-3 SHAP sensors, the bottom-3 SHAP sensors, and ten random draws of 3 sensors.
Cross-method test. Permutation importance [23], an attribution mechanism unrelated to Shapley values, is computed on the same model and validation set (10 repetitions per feature), and the agreement between the two rankings is quantified by the Spearman rank correlation over all 14 features.

J. Attribution instability and the Agreement Analysis

RQ3 requires a per-instance measure of how trustworthy an explanation is. Following the perturbation-robustness view of interpretability [9], the stability of an attribution at x is probed by jittering the input: K = 20 noisy copies x + εk are drawn from N(0, σ2IM), with σ = 0.02 on the unit-scaled features (clipped to [0, 1]), SHAP is recomputed for every copy, and the instability score is the dispersion of the attribution vector across copies,

$$I(x) = \frac{1}{M} \Sigma_j \, std_{k=1\ldots K}(\varphi_j(x + \varepsilon_k)) \qquad (13)$$

A small I(x) indicates that the explanation is stable at scales below the sensor noise floor. Conversely, a large I(x) indicates the stated reasons for the prediction are themselves unstable. Here we compute I(x) and the calibrated prediction interval width W(x) = q̂hi(x) − q̂lo(x) + 2Q on the same modeling pipeline (I explains the point GBR, while W comes from its conformalized quantile layer), using n = 300 validation points sampled uniformly at random and report Spearman and Pearson measures of their correlation. The tested hypothesis is positive correlation between W and I, such that the calibration-adjusted interval width serves as a warning flag for fragility of explanation. Either verification or failure of this hypothesis is instructive: if W and I are found to be independent quantities then uncertainty and explanation quality must be validated separately.

III. RESULTS

A. Reference Model Performance

Table 1 reports the test accuracies of both reference models on the 80/20 split. Metrics are evaluated on the official test set of 100 engines, using raw (uncapped) ground truth values. LSTM consistently outperforms GBR across metrics, improving test RMSE by 19.5% (from 18.21 to 14.65 cycles) and score S by 35.9% (from 575.8 to 369.3), which is consistent with the demonstrated edge of recurrent models over shallow learners on this benchmark [4]. Both fall within accepted variance of the FD001 literature and are therefore suitable seeds for the trustworthiness layer. Fig. 4 shows degradation trajectories of the four highest-ranked sensors for representative engines; Fig. 5 shows the GBR baseline on the validation engines, and Fig. 6 the LSTM training history.

Tab. 1. Reference model performance on FD001. Test metrics are computed on the official test set against uncapped ground truth.

Model	Validation RMSE	Test RMSE	Score S
Gradient boosting (GBR)	17.12	18.21	575.8
LSTM (window = 30)	11.71	14.65	369.3

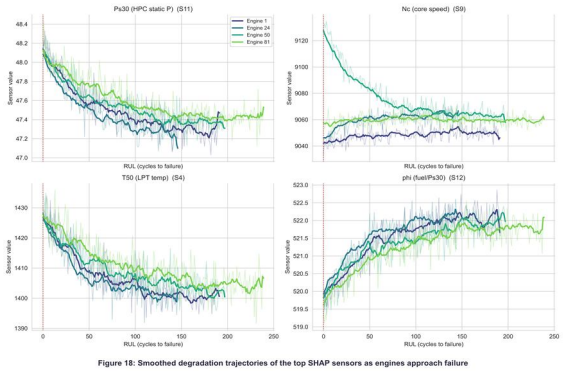


Fig. 4. Degradation trajectories of the four highest-ranked sensors (Ps30, Nc, T50, phi) for four representative engines, plotted against cycles to failure.

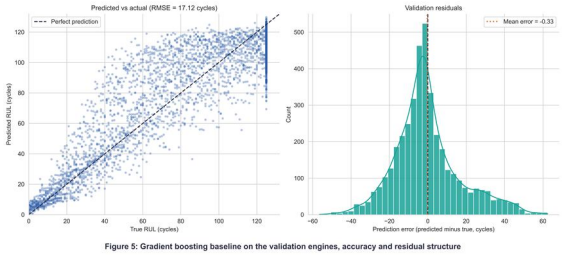


Fig. 5. GBR baseline on the validation engines: predicted versus actual RUL and residual structure (RMSE 17.12 cycles).

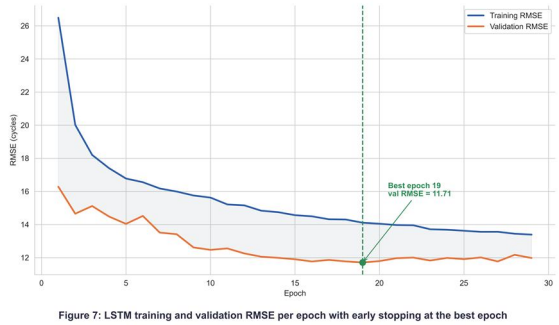


Fig. 6. LSTM training and validation RMSE per epoch; early stopping restores the best weights at epoch 19 (validation RMSE 11.71 cycles).

B. Conformal calibration (RQ1)

Table 2 shows overall calibration using the 60/20/20 split at the 90% nominal level. Fixed-width split conformal prediction undercovers on both models, achieving calibration values of 84.7% (GBR, width 60.3 cycles) and 86.2% (LSTM, width 42.0 cycles, calibrated half-width $\hat{q} = 21.0$). The marginal deficit itself lies within expected split-to-split variability: although the calibration set contains 4,262 rows, they cluster within only 20 engines, so the effective sample size behind Eq. (7) is small and the coverage realized on a single split can deviate from nominal by several percentage points. Table 3 disaggregates by life stage to diagnose where the calibration fails: at a fixed width, the GBR interval substantially overcovers in the near-failure regime (99.8% coverage for RUL 0–30, where late-life predictions are most accurate and the fixed width is much too generous) and just as substantially undercovers in the transition regime (74.5% for RUL 60–90, where an engine's transition away from healthy behavior is intrinsically difficult to predict). The LSTM shows the same pattern (100.0%, 84.7%, 68.5%, and 88.0% for those four bands), so this life-stage-dependent miscalibration is due to the heteroscedastic nature of the problem, rather than any particular model.

Tab. 2. Aggregate conformal calibration on held-out validation engines (nominal coverage 90%).

Method	Model	Coverage	Mean width
Split conformal (fixed width)	GBR	84.7%	60.3
CQR (adaptive width)	GBR	89.3%	56.1
Split conformal (fixed width)	LSTM	86.2%	42.0

CQR removes most of this defect. Aggregate coverage improves to 89.3% and average width shrinks to 56.1 cycles, a simultaneous improvement in calibration and precision. More strikingly, width adapts to the input: Table 3 shows that CQR allows itself 24.8 cycles of width near failures where the degradation signal was strong, and 69.5 cycles in the transition region where it was weak, changing by a factor of 2.8. Band-wise coverages narrow to 79.5–92.8%, including precisely 90.5% in the band where the fixed-width method failed. The redistribution of width comes at the one clear trade-off we report: coverage in the near-failure band, previously overcovered at 99.8%, comes down to 79.5%, modestly below nominal.

Fig. 7 shows the resulting per-engine intervals on the test set, sorted by true RUL. Fig. 8 confirms calibration at each nominal level from 50% to 95%. At these levels the empirical coverage tracks the requested coverage to within 1.9 percentage points everywhere (50.5, 59.3, 69.1, 78.1, 89.3, and 94.3%, respectively). Fig. 9 summarizes how CQR restores even band-wise coverage by adapting interval width to local prediction difficulty.

Tab. 3. Coverage and width stratified by RUL band (GBR, validation engines, nominal 90%).

RUL band	n	Coverage, fixed	Coverage, CQR	Width, CQR
0–30	600	99.8%	79.5%	24.8
30–60	600	83.0%	84.7%	58.4
60–90	600	74.5%	90.5%	69.5
90–125	2299	83.8%	92.8%	60.2

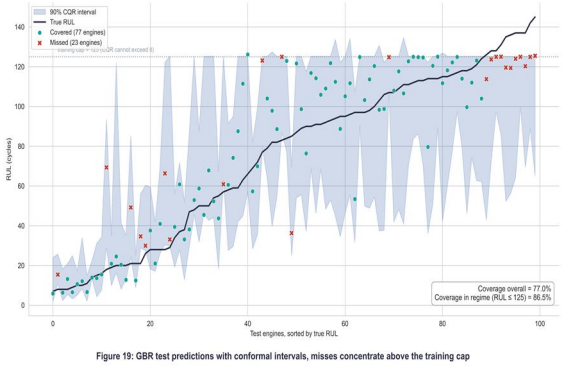


Fig. 7. Predicted versus actual RUL with 90% CQR intervals for the 100 test engines, sorted by true RUL.

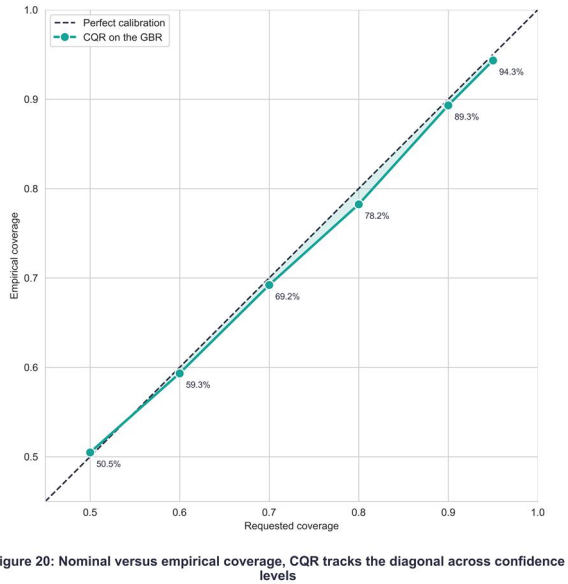


Fig. 8. Calibration plot comparing empirical vs. nominal coverage of CQR intervals across six coverage levels.

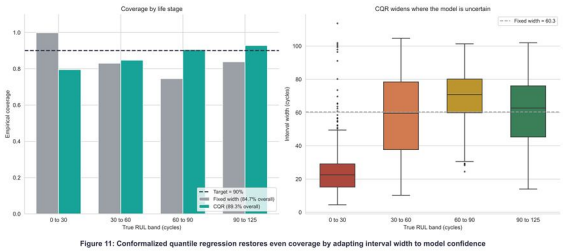


Fig. 9. CQR restores even band-wise coverage by adapting interval width to local prediction difficulty.

On the officially supplied test set, for which ground truth is uncapped, total coverage is 77.0%. The majority of the coverage deficit can be explained by the 11 of 100 engines whose true RUL is beyond the training-set cap of 125 cycles; the conditional coverage gap on these engines is by construction nonzero, as a cap-trained model cannot in general provide coverage on them. Coverage conditioned on remaining in the operating regime $y \leq R_{max}$, where maintenance decisions are actually being made, is 86.5% across the remaining 89 engines. This distinction between covered and uncovered engines is directly implied by the piecewise-linear labeling convention, and is mentioned here explicitly so that the above calibration result is understood within its appropriate context of validity.

C. Faithfulness of SHAP attributions (RQ2)

All analyses in this and the following subsection use the GBR trained on the 60/20/20 conformal split, whose uncorrupted validation RMSE is 20.70 cycles (Table 5). Table 4 presents the global SHAP importance ranking on this GBR, paired with permutation importance. At the top of the ranking are physically coherent indicators of high-pressure-compressor degradation, the simulated fault mode of FD001: static pressure at HPC outlet (Ps30; sensor 11), physical core speed (Nc; sensor 9), fuel-flow ratio phi (sensor 12), LPT outlet temperature (T50; sensor 4), and HPC outlet pressure (P30; sensor 7). The two rankings correlate at Spearman $\rho = 0.868$ ($p = 5.7 \times 10{-}5$), with the same six sensors appearing in the top six of both.

Tab. 4. Global importance rankings: mean |SHAP| versus permutation importance (GBR, validation sample).

Sensor	Physical quantity	Mean |SHAP|	Perm. importance
11	Ps30, HPC outlet static pressure	8.088	2.210
9	Nc, physical core speed	5.982	3.693
12	phi, fuel flow / Ps30	5.048	1.219
4	T50, LPT outlet temperature	4.222	0.691
7	P30, HPC outlet pressure	3.842	0.898
14	NRc, corrected core speed	3.109	1.067
15	BPR, bypass ratio	2.844	0.392
20	W31, HPT coolant bleed	2.417	0.204
2	T24, LPC outlet temperature	1.894	0.121
13	NRf, corrected fan speed	1.736	0.405
21	W32, LPT coolant bleed	1.723	0.155
17	htBleed, bleed enthalpy	1.540	0.094
3	T30, HPC outlet temperature	1.390	0.028
8	Nf, physical fan speed	1.022	0.336

The perturbation test in Table 5 verifies that this ranking is causal with respect to the model rather than merely descriptive. When we permute the three highest-ranked sensors, validation RMSE increases by 8.22 cycles. When we permute three randomly chosen sensors, validation RMSE increases by 2.69 cycles on average. When we permute the three lowest-ranked sensors, validation RMSE increases by 0.65 cycles. This 12.6-fold separation between the highest and lowest rankings is exactly the monotone ordering that a faithful attribution must produce. Combined with the cross-method agreement, we can answer RQ2 affirmatively: for this model, SHAP correctly identifies which sensors the prediction computation depends on. Fig. 10 shows the corresponding beeswarm summary plot. The direction of each sensor's effect agrees with the physics of compressor degradation. Fig. 11 visualizes the audit.

Tab. 5. Perturbation-based faithfulness audit (GBR, validation set; random condition averaged over 10 draws).

Condition	RMSE (cycles)	ΔRMSE
No corruption (baseline)	20.70	—
Top-3 SHAP sensors permuted	28.92	+8.22
Random 3 sensors permuted (mean)	23.39	+2.69
Bottom-3 SHAP sensors permuted	21.35	+0.65

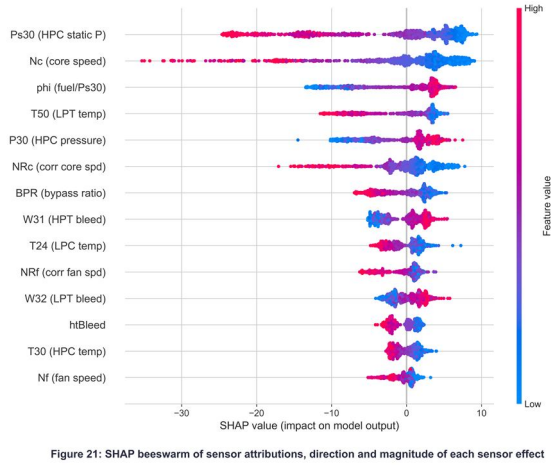


Fig. 10. SHAP beeswarm summary over 1,000 sampled validation predictions, features ordered by global importance.

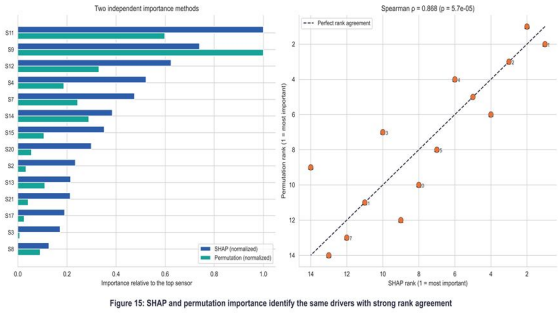


Fig. 11. Perturbation faithfulness audit: corrupting the top-ranked sensors damages accuracy far more than corrupting the lowest-ranked sensors.

D. Uncertainty–explanation agreement (RQ3)

This paper's key finding concerns the relationship between the two measures of trust calculated on the same modeling pipeline. Across the n = 300 validation points we sampled, CQR widths range from 8.2 to 101.0 cycles, instability ranges from 0.126 to 1.607, and the two quantities are strongly and positively correlated: Spearman ρ = 0.733 (p = 7.8 × 10−52) and Pearson r = 0.709 (p = 4.4 × 10−47). Calibration interval width is therefore associated with explanation instability, meaning that predictions made with tight calibrated intervals also have stable explanations, while predictions made with loose intervals have explanations that change drastically under perturbations below the sensor noise floor. Fig. 12 shows this relationship with points colored by true RUL: the region of narrow-and-stable predictions concentrates near failure where the degradation signature is unambiguous, while wide-and-unstable predictions concentrate far from failure, in the healthy and transition regimes. This is consistent with prediction difficulty acting as a common cause of both effects. Fig. 13 traces instability across the engine life: it rises far from failure, where prediction is hardest.

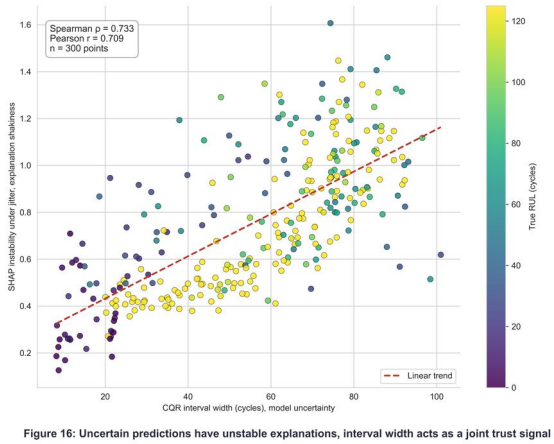


Fig. 12. CQR interval width versus SHAP instability for 300 validation points, colored by true RUL (Spearman ρ = 0.73). The positive association is the central empirical finding.

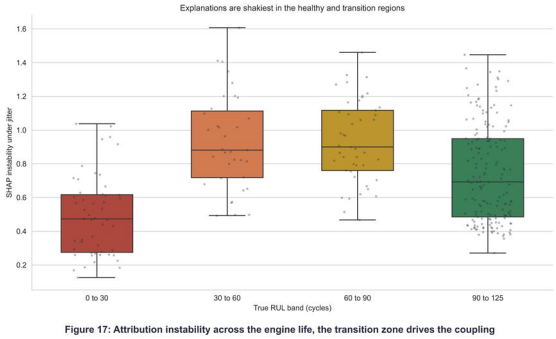


Fig. 13. Attribution instability across the engine life; instability rises far from failure where prediction is hardest.

IV. DISCUSSION

Together, these three results yield a concrete deployment recipe. RQ1 certifies that the interval width is calibrated, so that it is a well-defined statistic rather than a rule-of-thumb confidence. RQ2 certifies that the explanations are faithful on average, so that variations in the instability metric reflect the model's actual decision process rather than artifacts of the attribution method. RQ3 then shows that the calibrated width quantifies that instability. The upshot for deployment practitioners is that a prognostics system only needs to report one scalar per prediction: when the 90% interval for a machine is small, both the predicted RUL and the sensor-level explanation provided to the user may be trusted; when it is large, both should be scrutinized. The interval width inherits the distribution-free validity of conformal prediction, while its role as a proxy for explanation instability is established empirically; both are obtained entirely by post-processing, at zero retraining cost, on top of whatever model an operator already has in production.

It is helpful to state a candidate mechanism for why the two quantities correlate. The correlation is taken to be evidence of a common cause, which is posited to be local prediction difficulty. In the early-warning region near the onset of degradation, the conditional distribution of RUL given the sensors is inherently dispersed, which causes CQR to produce wide intervals, and at the same time the learned regression surface is steep and jagged locally, which causes nearby input perturbations to produce volatile Shapley values. We do not claim that either quantity causes or explains the other; we claim only that they co-vary consistently, and that one can be used to predict the other, which is sufficient for the operational recommendation above.

The calibration results come with two caveats. First, the redistribution of width achieved by CQR leaves the near-failure band slightly undercovered (79.5%), so the adaptive procedure is a partial fix, not a full correction; methods that target group-conditional coverage, such as Mondrian conformal prediction, are a natural extension. Second, the cap-regime discussion of Section 3.2 illustrates that the coverage guarantee of Eq. (7) must be understood in light of the label convention adopted by the benchmark: a model trained to predict capped labels is by construction incapable of conformally covering any engine whose true RUL exceeds the cap; reporting only the conditional 86.5% statistic would therefore overstate the method's success, just as reporting only the aggregate 77.0% would overstate its failure.

This work has focused exclusively on FD001, the single-condition single-fault subset of C-MAPSS. Confirming that the sensor ranking, the band-wise calibration pattern, and most importantly the width–instability correlation hold on FD002–FD004, datasets with multiple operating regimes and fault modes, is the obvious next step and would provide a more stringent validation of the trust-signal hypothesis. The agreement analysis is computed only on the GBR for which exact SHAP values can be computed; porting it to the LSTM would require approximate explainers (DeepSHAP or kernel-based) as well as quantile-loss training for the adaptive intervals, and would show whether the correlation generalizes across architectures. The measure of instability from Eq. (13) relies on the full attribution vector; rank-based variants that consider only the identity of the top sensors may correspond more closely to how engineers use explanations. In addition, all conformal results are reported for a single engine-level split; averaging coverage over repeated random splits would tighten the empirical statements of Section 3.2. Lastly, the jitter scale σ was set to 0.02; analyzing how the instability metric changes with σ, and comparing across different attribution methods, would strengthen the construct validity of our proposed metric.

V. CONCLUSIONS

This paper cast RUL prediction on C-MAPSS as a problem of prediction trustworthiness rather than accuracy alone, and provided the previously unavailable link between its two components. Conformalized quantile regression was demonstrated to provide near-nominal, locally adaptive prediction intervals on a leakage-controlled FD001 benchmark, relative to fixed-width conformal prediction, which was substantially miscalibrated across life stages; a two-part audit showed SHAP values on this dataset to be faithful, with a 12.6-fold difference in perturbation effect between highest- and lowest-ranked sensors and a rank correlation of $\rho = 0.868$ between methods; finally, this calibrated interval width was shown to correlate strongly with the variation in those attributions ($\rho = 0.733$, $p = 7.8 \times 10{-}52$) when tested against input jittering. Not only do calibrated uncertainty estimates and feature attributions agree, but their agreement endows the width of the conformal prediction interval with utility as a single, distribution-free, retraining-free trust metric for safety-critical mechanical system maintenance.

REFERENCES

[1] Saxena A, Goebel K, Simon D, Eklund N. Damage propagation modeling for aircraft engine run-to-failure simulation. In: Proceedings of the International Conference on Prognostics and Health Management; Denver, CO. 2008. p. 1-9. https://doi.org/10.1109/PHM.2008.4711414

[2] Frederick D, DeCastro J, Litt J. User's guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS). NASA Technical Report TM-2007-215026. Cleveland, OH: NASA Glenn Research Center; 2007.

[3] Heimes FO. Recurrent neural networks for remaining useful life estimation. In: Proceedings of the International Conference on Prognostics and Health Management; Denver, CO. 2008. p. 1-6. https://doi.org/10.1109/PHM.2008.4711422

[4] Zheng S, Ristovski K, Farahat A, Gupta C. Long short-term memory network for remaining useful life estimation. In: Proceedings of the IEEE International Conference on Prognostics and Health Management (ICPHM); 2017. p. 88-95. https://doi.org/10.1109/ICPHM.2017.7998311

[5] Hu C, Youn BD, Wang P, Yoon JT. Ensemble of data-driven prognostic algorithms for robust prediction of remaining useful life. Reliability Engineering & System Safety 2012; 103: 120-135. https://doi.org/10.1016/j.ress.2012.03.008

[6] Li X, Ding Q, Sun JQ. Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety 2018; 172: 1-11. https://doi.org/10.1016/j.ress.2017.11.021

[7] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30; 2017. p. 4765-4774.

[8] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2020; 2(1): 56-67. https://doi.org/10.1038/s42256-019-0138-9

[9] Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. In: Proceedings of the ICML Workshop on Human Interpretability in Machine Learning; 2018.
[10] Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems 31; 2018. p. 9505-9515.
[11] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020. p. 180-186. https://doi.org/10.1145/3375627.3375830
[12] Baptista ML, Goebel K, Henriques EMP. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. Artificial Intelligence 2022; 306: 103667. https://doi.org/10.1016/j.artint.2022.103667
[13] Vovk V, Gammerman A, Shafer G. Algorithmic Learning in a Random World. New York, NY: Springer; 2005.
[14] Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. Journal of the American Statistical Association 2018; 113(523): 1094-1111. https://doi.org/10.1080/01621459.2017.1307116
[15] Angelopoulos AN, Bates S. Conformal prediction: a gentle introduction. Foundations and Trends in Machine Learning 2023; 16(4): 494-591. https://doi.org/10.1561/2200000101
[16] Romano Y, Patterson E, Candès E. Conformalized quantile regression. In: Advances in Neural Information Processing Systems 32; 2019. p. 3543-3553.
[17] Javanmardi A, Hüllermeier E. Conformal prediction intervals for remaining useful lifetime estimation. International Journal of Prognostics and Health Management 2023; 14(2). https://doi.org/10.36001/ijphm.2023.v14i2.3417
[18] Cordier T, Blot V, Lacombe L, Morzadec T, Capitaine A, Brunel N. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. In: Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications, Proceedings of Machine Learning Research 204; 2023. p. 549-581.
[19] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001; 29(5): 1189-1232. https://doi.org/10.1214/aos/1013203451
[20] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation 1997; 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
[21] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011; 12: 2825-2830.
[22] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI); Savannah, GA. 2016. p. 265-283.
[23] Breiman L. Random forests. Machine Learning 2001; 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

Conflicts of interest

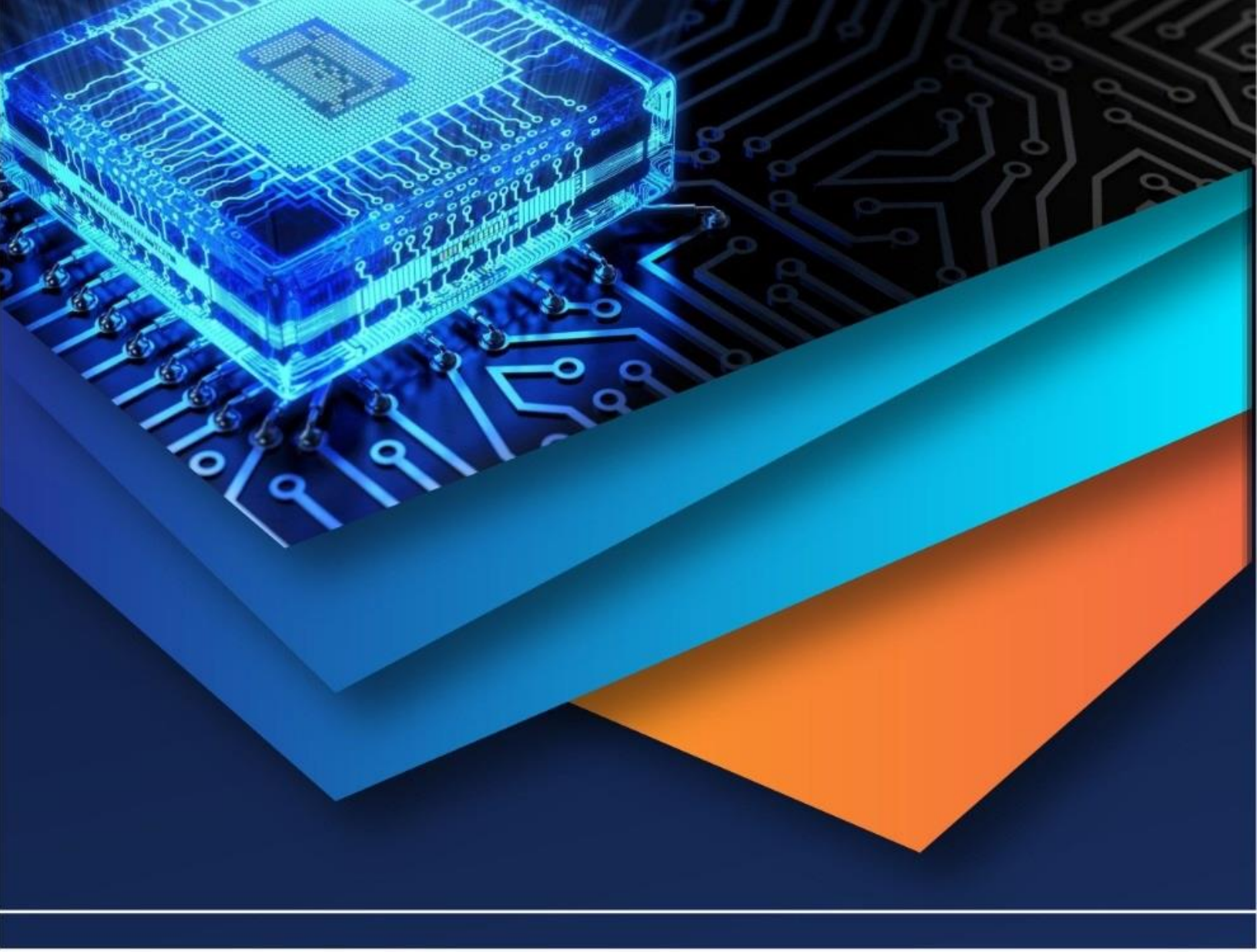
The authors declare no conflicts of interest.

Data availability statement

The NASA C-MAPSS dataset analysed in this study is publicly available from the NASA Prognostics Center of Excellence data repository.

Acknowledgements

The authors thank NASA's Prognostics Center of Excellence for making the C-MAPSS dataset publicly available. The authors used Claude (Anthropic) to assist with language editing and manuscript formatting.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY