Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 02, Jul-Dec 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.62.106

Research Paper

Calibrated but not safe: Conformal prediction intervals for turbofan remaining useful life near end of life and under operating shift

Nuno Alexandre Soares Domingues[1*], Danish Ather[2]
[1*]Adjunct Professor, Department of Mechanical Engineering, Instituto Superior de Engenharia de Lisboa (ISEL), Instituto Politécnico de Lisboa, Lisbon, Portugal.
[2]Professor, Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan.

Article Info

Article History:
Received: 27 June 2026
Revised: 26 August 2026
Accepted: 16 September 2026
Published: 22 October 2026

Keywords:
Remaining Useful Life
Conformal Prediction
Uncertainty Quantification
Prognostics and Health Management
Safety-Critical Systems
C-MAPSS

ABSTRACT

Machine learning models that predict the remaining useful life (RUL) of aircraft engines are increasingly paired with conformal prediction, which converts point predictions into intervals with a guaranteed average coverage. In a safety-critical maintenance setting, however, an average guarantee is not sufficient: what matters is whether the interval is reliable and informative for engines that are close to failure, and whether it remains valid when the fleet differs from the data used for calibration. This study evaluates five interval methods built on ridge regression, random forest, gradient boosting and multilayer-perceptron predictors using the NASA C-MAPSS turbofan benchmark (FD001–FD004, 707 test engines), with repeated engine-level calibration splits. Coverage, width and the rate at which the interval overestimated the true RUL were measured overall and for engines with a true RUL of 30 cycles or less. Global split conformal prediction met the nominal 90% coverage on average (89–92%) and covered almost every engine near failure, but with intervals about 44 cycles wide, which is uninformative when only 30 cycles remain; it also undercovered engines with 61–90 cycles of remaining life (68–76%). Calibrating separately within bands of predicted RUL (Mondrian conformal prediction) halved the width near failure for tree ensembles (from 43.7 to 20.7 cycles for gradient boosting, $p = 3 \times 10-8$) while keeping near-failure coverage at 96.5–97.8%, but failed for ridge regression (77%). Uncalibrated quantile regression overestimated the remaining life of 20% of near-failure engines, and conformalised quantile regression restored average coverage, but its near-failure coverage fell as low as 80% and its near-failure intervals averaged over 90 cycles. When models calibrated on FD001 were applied to engines with an additional fault mode (FD003), near-failure coverage collapsed to 37–50%, and 95% of all miscoverage under shift was overestimation of remaining life, the unsafe direction. Conformal guarantees are therefore marginal and conditional on

Check for updates

Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN)

exchangeability; prognostic intervals should be evaluated near end of life and under realistic shift before they are used for maintenance decisions.

Corresponding Author:
Nuno Alexandre Soares Domingues
Adjunct Professor, Department of Mechanical Engineering, Instituto Superior de Engenharia de Lisboa (ISEL), Instituto Politécnico de Lisboa, Lisbon, Portugal.
Email: nuno.domingues@isel.pt

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Predicting the remaining useful life (RUL) of a component, that is, the number of operating cycles before it fails, is central to prognostics and health management and to condition-based maintenance of safety-critical assets such as aircraft engines [1], [2]. Since the release of the NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) turbofan data [3], [4], data-driven RUL models have progressed from similarity-based methods [5] and recurrent networks [6] to convolutional [7], [8] and long short-term memory networks [9], and benchmark errors have fallen steadily [10]. A point estimate of RUL, however, does not tell a maintenance planner how much to trust it. For safety-critical decisions, the uncertainty of the estimate matters as much as the estimate itself [11], [12].

Conformal prediction offers an attractive solution because it turns any point predictor into a prediction interval with a finite-sample, distribution-free coverage guarantee, provided that calibration and test data are exchangeable [13], [14], [15], [16]. Javanmardi and Hüllermeier [17] showed that conformal intervals for C-MAPSS achieve their nominal coverage, and Robinson [18] recently proposed asymmetric conformalised quantile regression to avoid late maintenance. The coverage guarantee, however, is marginal: it holds on average over all test engines, not for particular groups [15], [19]. In prognostics, the group that matters most is the engines close to failure, for which an interval that is too wide is useless and an interval that overestimates the remaining life is dangerous. Moreover, the guarantee depends on exchangeability, which is violated when a model calibrated on one fleet is applied to engines with different operating conditions or failure modes [20], [21].

This paper examines how conformal RUL intervals behave in exactly these two respects. The contributions are as follows:

- A controlled evaluation of five interval methods (global split conformal prediction with two calibration designs, Mondrian conformal prediction by predicted-RUL band, uncalibrated quantile regression and conformalised quantile regression) on four point predictors and all four C-MAPSS sub-datasets, with repeated engine-level calibration splits.
- Evidence that global conformal intervals meet their nominal average coverage but are too wide to be informative near end of life, and that they undercover engines in the transition from healthy to degrading operation.
- Evidence that band-wise (Mondrian) calibration halves the interval width for near-failure engines with tree-ensemble predictors without sacrificing their coverage, but fails when the predictor is poorly calibrated near end of life.
- A quantification of failure under operating and fault-mode shift, showing that near-failure coverage collapses and that almost all miscoverage is in the unsafe direction of overestimating remaining life.

2. RELATED WORK

2.1 Data-Driven Rul Prediction

[1], [2] review statistical and machine learning approaches to RUL estimation. On C-MAPSS, Heimes [6] used recurrent neural networks [5] a similarity-based approach. [7], [8] convolutional networks, [9] long short-term memory networks. A common modelling convention, which is also adopted here, caps the RUL target at a constant in the early, healthy phase of operation, because degradation is not observable there [6], [8]. Ramasso and Saxena [10] benchmarked prognostic methods on C-MAPSS and emphasised the need for consistent evaluation.

2.2 Uncertainty Quantification and Conformal Prediction

Uncertainty in neural predictors is often estimated with Monte Carlo dropout [22] or deep ensembles [23], but these estimates carry no coverage guarantee [12]. Conformal prediction [13], in its split or inductive form [14], [15], calibrates a nonconformity score on held-out data and guarantees that the resulting intervals contain the true value with at least the nominal probability when calibration and test points are exchangeable. Conformalised quantile regression (CQR) [24] combines this guarantee with the adaptive width of quantile regression [25]. Vovk [19] showed that marginal validity does not imply validity within subgroups and proposed Mondrian (category-wise) calibration to obtain group-conditional coverage. [20], [21] Extended conformal prediction to covariate shift and non-exchangeable data, respectively, and Angelopoulos and Bates [16] provide an accessible introduction.

2.3 Conformal Prediction for RUL

Javanmardi and Hüllermeier [17] applied three conformal algorithms to convolutional and gradient-boosting RUL predictors on C-MAPSS and reported valid marginal coverage. Robinson [18] used an asymmetric formulation of CQR to favour conservative intervals and reported coverage of 88–92%. Neither study examined coverage specifically for engines close to failure or the effect of shifting to sub-datasets with different fault modes or operating conditions, which are the focus here.

3. METHODOLOGY

Figure 1 summarises the experimental pipeline. All splits were made at the level of whole engines, so that no engine contributed data to more than one of the training, calibration and test sets.

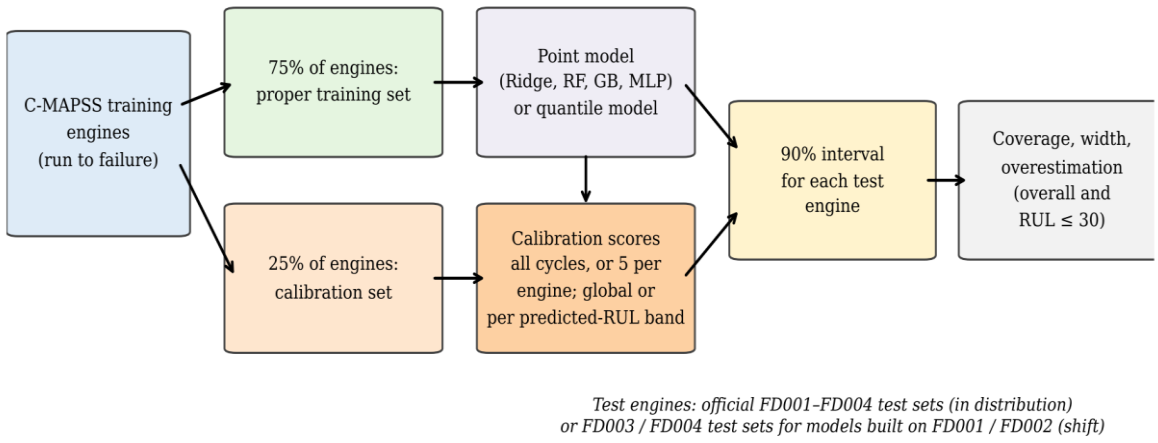


Figure 1. Experimental Pipeline. Training Engines are Split into a Proper Training Set and a Calibration Set; Point or Quantile Models are Fitted on the Former and Calibrated on the Latter, and the Resulting 90% Intervals are Evaluated on the Official Test Engines

3.1 Data

The C-MAPSS data [3], [4] simulate the degradation of a fleet of turbofan engines from an unknown initial wear state to failure. Each sub-dataset contains run-to-failure trajectories for training and truncated trajectories for testing, together with the true RUL at the end of each test trajectory Table 1. FD001 and FD003 were recorded at a single operating condition, and FD002 and FD004 at six operating conditions; FD001 and FD002 contain one fault mode (high-pressure compressor degradation), and FD003 and FD004 two (high-pressure compressor and fan degradation). Each record contains three operating settings and 21 sensor measurements per cycle. Following common practice, the 14 sensors that vary with degradation were used Figure 2.

Table 1. The C-Mapss Sub-Datasets

Sub-Dataset	Operating Conditions	Fault Modes	Training Engines	Test Engines	Test Engines with True RUL ≤ 30	Calibration Engines per Split
FD001	1	1	100	100	25	25
FD002	6	1	260	259	61	65
FD003	1	2	100	100	20	25
FD004	6	2	249	248	53	62

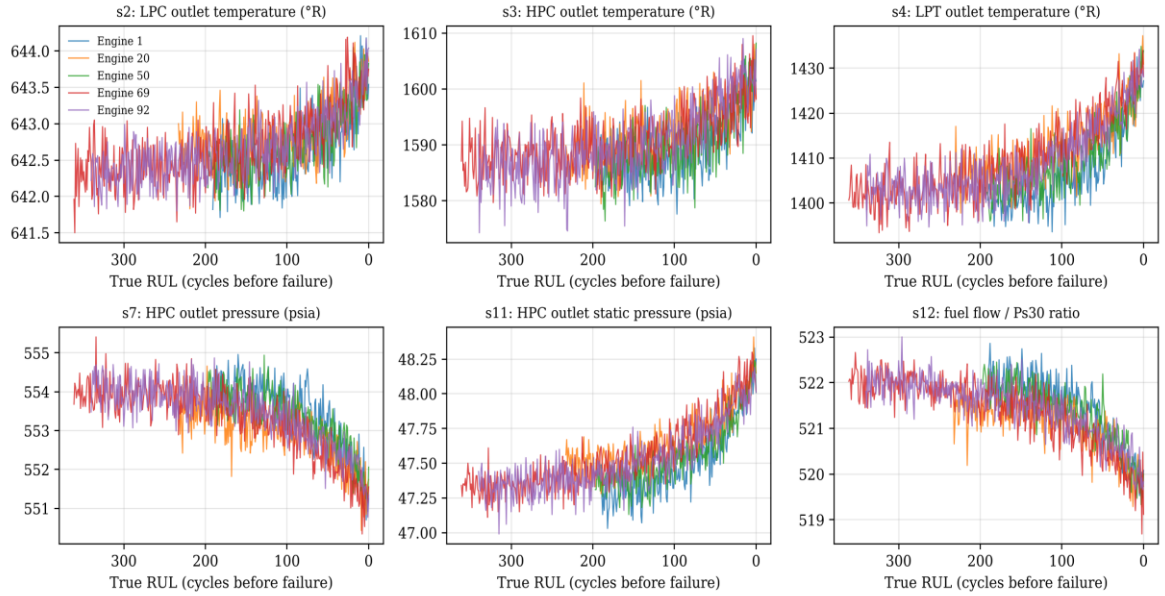


Figure 2. Six of the Fourteen Sensor Signals Used in the Study For Five Real FD001 Training Engines, Plotted Against the True Remaining Useful Life. Degradation Becomes Visible in the Last 100–150 Cycles before Failure

3.2 Pre-Processing, Features and Target

Sensor values were standardised using the mean and standard deviation of the training data; for FD002 and FD004, standardisation was performed separately within each of the six operating conditions, identified by k-means clustering of the operating settings. For every cycle with at least 30 cycles of history, a feature vector was computed from the preceding 30-cycle window: the mean of the last five cycles, the window mean, the window standard deviation and the least-squares slope of each sensor, together with the number of cycles observed, giving 57 features. Test engines were represented by the window ending at their last observed cycle (shorter windows were used for the few test engines with fewer than 30 cycles). The target was the piecewise-linear RUL capped at 125 cycles [6], [8]:

[1] Where Tfail is the failure cycle of the engine. The same cap was applied to the true RUL of test engines when computing interval coverage, which is equivalent to treating any RUL above 125 cycles as "healthy".

3.3 Point and Quantile Predictors

Four point predictors were trained on the proper training set with scikit-learn [26]: ridge regression [27] on standardised features ($\alpha = 1$); a random forest [28] with 60 trees, 30% of the features considered at each split and at least five samples per leaf; histogram gradient boosting (GB) [29] with 300 iterations and learning rate 0.05; and a multilayer perceptron (MLP) [30] with two hidden layers of 64 and 32 units, trained with Adam [31] and early stopping. For the quantile-based methods, two LightGBM models [32] with the pinball loss [25] were trained to predict the 5% and 95% conditional quantiles.

3.4 Interval Methods

All intervals targeted 90% coverage ($\alpha = 0.1$). In each of several random splits, 25% of the training engines were assigned to the calibration set. Let $\hat{y}(x)$ denote the point prediction and let $s_i = |y_i - \hat{y}(x_i)|$ be the absolute residual of calibration point i. Split conformal prediction [14], [15] forms the interval $\hat{y}(x) \pm \hat{q}$, where $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil$-th smallest of the n calibration scores: [2]

Two calibration designs were compared. In the all-cycles design, every window of every calibration engine is a calibration point. In the engine-subsampled design, five windows are drawn at random from each calibration engine, which reduces the dominance of long-lived engines and the dependence between points from the same engine. Mondrian conformal prediction [19] applies the same quantile separately within three bands of predicted RUL (below 50, 50–100 and above 100 cycles), using engine-subsampled calibration points, so that engines predicted to be near failure receive an interval calibrated only on similar engines: [3]

Where $b(\cdot)$ maps a prediction to its band. Quantile regression uses the predicted 5% and 95% quantiles $[q_{lo}(x), q_{hi}(x)]$ directly, without calibration. Conformalised quantile regression (CQR) [24] widens these quantiles by the conformal quantile of the score [4]

Computed on the engine-subsampled calibration points.

3.5 Evaluation

Point predictors were evaluated on the official test sets by the root-mean-square error (RMSE) and the asymmetric C-MAPSS scoring function [3], which penalises late predictions (overestimates of RUL) more heavily than early ones: [5]

where d is the prediction minus the true RUL. Intervals were evaluated by their coverage (the proportion of test engines whose true RUL lies in the interval), their mean width, and the proportion of engines whose true RUL lies below the interval, which corresponds to an overestimate of remaining life and is the unsafe direction of miscoverage. All quantities were computed for all test engines and for the 159 near-failure engines with a true RUL of at most 30 cycles Table 1. Coverage was also broken down into four bands of true RUL (0–30, 31–60, 61–90 and 91–125 cycles).

3.6 Distribution Shift

To examine shift, models trained and calibrated on FD001 were applied to the FD003 test engines, which add a second fault mode, and models trained and calibrated on FD002 were applied to the FD004 test engines, which add the same fault mode under six operating conditions. In both cases the operating conditions of source and target agree, so the shift is caused by the additional failure mechanism.

3.7 Repetitions and Statistics

Each configuration was repeated with eight random calibration splits for FD001, FD003 and FD001 → FD003, and five for FD002, FD004 and FD002 → FD004, the smaller number reflecting the larger computational cost of the multi-condition data. Results are reported as means over splits with standard deviations where relevant. Paired comparisons between interval methods used the Wilcoxon signed-rank test [33] over the 26 in-distribution runs (splits × sub-datasets). All computations were carried out in Python with scikit-learn 1.8 and LightGBM on a two-core CPU.

4. RESULTS AND DISCUSSION

4.1 Point Prediction Accuracy

Table 2 reports the accuracy of the four point predictors trained on 75% of the training engines. Gradient boosting and the random forest were the most accurate, with RMSE of 11.9–14.4 cycles, which is in the range reported for convolutional and recurrent networks on these data [8], [9]. Ridge regression and the MLP were 2–5 cycles less accurate. Under fault-mode shift, the tree ensembles degraded to RMSE of 19.5–21.5, whereas ridge regression and the MLP extrapolated to implausible values for some engines, with RMSE of 62–190 cycles; their scoring-function values, dominated by a few extreme late predictions, are therefore not reported.

Table 2. Point Prediction Accuracy on the Official Test Sets: Rmse and C-Mapss Score (Mean ± Sd Over Calibration Splits; Lower Is Better)

Model	FD001 RMSE	FD002 RMSE	FD003 RMSE	FD004 RMSE	FD001 Score	FD004 Score	FD001 → FD003 RMSE	FD002 → FD004 RMSE
Ridge	16.2 ± 0.2	16.6 ± 0.1	16.5 ± 0.2	19.2 ± 0.1	411 ± 7	1,877 ± 35	156.2 ± 30.7	61.9 ± 8.4
Random forest	12.0 ± 0.3	13.1 ± 0.1	13.1 ± 0.5	14.4 ± 0.2	214 ± 20	1,010 ± 55	19.6 ± 0.7	19.5 ± 0.3
Gradient boosting	11.9 ± 0.4	13.0 ± 0.3	12.6 ± 0.7	14.2 ± 0.1	214 ± 24	1,007 ± 67	21.5 ± 0.7	21.4 ± 1.5
MLP	14.9 ± 0.8	16.3 ± 0.7	15.8 ± 0.6	16.2 ± 1.1	360 ± 54	1,784 ± 490	189.5 ± 35.9	166.2 ± 29.2

4.2 Average Coverage Is Met, but Not Where It Matters

Table 3 summarises the interval methods over the four in-distribution sub-datasets. Global split conformal prediction met the nominal level on average, with coverage of 88.9–90.5% for the all-cycles design and 90.5–91.9% for the engine-subsampled design. Engine subsampling raised coverage by 1.5 percentage points on average ($p = 1.4 \times 10-9$ overall model–dataset–split combinations) at the cost of intervals about 3 cycles wider. Near failure, the global intervals covered virtually every engine (94.5–99.8%), but they were as wide there as everywhere else, 43.7–59.4 cycles for an engine with at most 30 cycles left. An interval that states that an engine with 15 cycles remaining has between −7 and 37 cycles left offers little guidance for scheduling maintenance.

Table 3. In-Distribution Interval Performance Averaged Over Fd001–Fd004 (90% Nominal Coverage). Coverage and Overestimation in %, Widths in Cycles

Predictor	Method	Coverage, All Engines	Coverage, RUL ≤ 30	Overestimated, RUL ≤ 30	Width, All Engines	Width, RUL ≤ 30
Gradient boosting	Split CP (all cycles)	89.2	99.6	0.4	43.7	43.7
	Split CP (engine-subsampled)	90.6	99.8	0.2	47.1	47.1
	Mondrian CP	92.1	97.8	2.2	44.5	20.7
Random forest	Split CP (all cycles)	89.2	99.6	0.4	44.7	44.7

	Split CP (engine-subsampled)	90.5	99.8	0.2	47.7	47.7
	Mondrian CP	91.7	96.5	3.5	44.6	22.3
MLP	Split CP (all cycles)	88.9	98.6	1.3	51.9	51.9
	Mondrian CP	89.8	92.2	5.7	52.5	30.6
Ridge	Split CP (all cycles)	90.5	94.5	2.3	56.7	56.7
	Mondrian CP	88.2	77.3	19.3	52.4	31.9
LightGBM quantiles	Quantile regression (uncalibrated)	85.4	78.1	20.3	57.0	85.4
	CQR	91.2	90.4	9.1	62.5	90.9

The breakdown by true-RUL band Figure 3 reveals a second problem that the average hides. All methods undercovered engines with 61–90 cycles remaining: global split conformal prediction covered only 68.3% of them on FD001 and 76.2% on FD002, while overcovering engines near failure. This band corresponds to the transition from the capped, healthy phase to visible degradation, where engines with similar sensor readings can be at quite different points of their life, so residuals are largest. A single global interval width is too large near failure and too small in the transition.

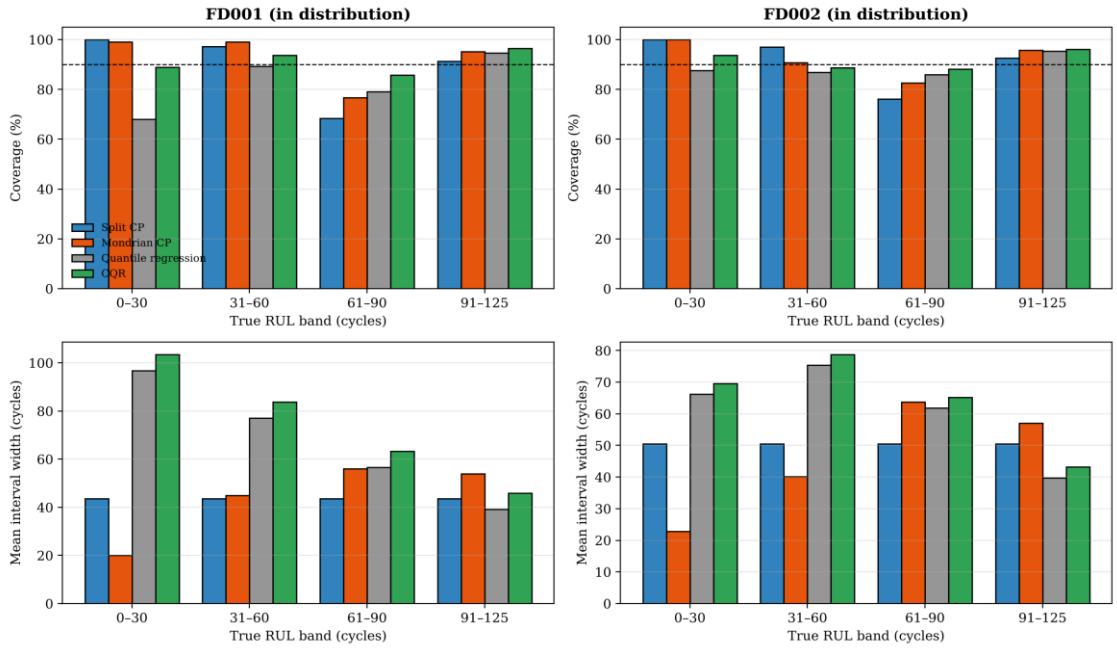


Figure 3. Coverage (Top) and Mean Width (Bottom) of 90% Intervals From Gradient Boosting (Split and Mondrian Cp) and Lightgbm Quantile Models (Quantile Regression and Cqr) by Band of True Rul, for Fd001 (Left; Eight Splits) and Fd002 (Right; Five Splits). The Dashed Line Marks the Nominal 90% Coverage

4.3 Band-Wise Calibration Narrows Near-Failure Intervals

Mondrian calibration within bands of predicted RUL addressed the width problem for the tree ensembles. Pooled over the 26 in-distribution runs, it reduced the near-failure width of gradient-boosting intervals from 43.7 to 20.7 cycles and that of random-forest intervals from 44.7 to 22.3 cycles (Wilcoxon p = 3 × 10−8 for both), while overall width was unchanged (p ≥ 0.62). Near-failure coverage fell slightly but

remained above nominal (97.8% and 96.5%, p = 0.018 and 0.001), and overall coverage rose to 91.7–92.1%. Figure 4 shows the effect on three real FD001 test engines: the band-wise interval tightens as the predicted RUL falls, whereas the lower bound of the global interval drops below zero cycles near failure.

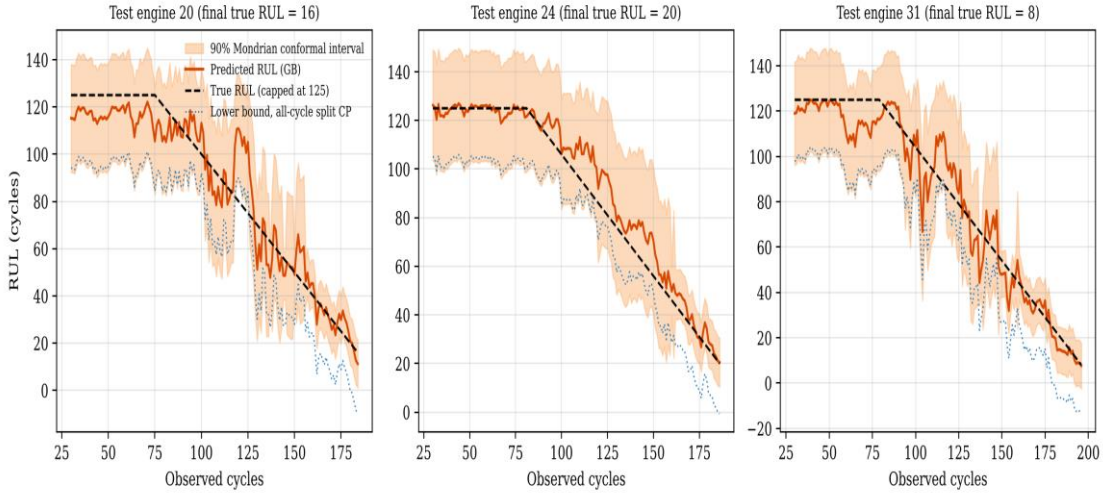


Figure 4. Gradient-Boosting Rul Predictions With 90% Mondrian Conformal Intervals over the Observed Life of Three Real Fd001 Test Engines that end Close to Failure. The Dotted Line Shows the Lower Bound of the Global Split-Conformal Interval, which Becomes Negative near the End tf the Trajectory

Mondrian calibration, however, conditions on the predicted rather than the true RUL, so it can only be as good as the predictor's ranking of engines. For ridge regression, whose near-failure predictions were less accurate, engines close to failure were frequently predicted to lie in a healthier band and received an interval calibrated for healthier engines. Its near-failure coverage fell to 77.3% and the remaining life of 19.3% of near-failure engines was overestimated, compared with 2.3% for the global interval Table 3. The MLP showed the same tendency to a lesser degree (92.2% coverage, 5.7% overestimated).

4.4 Quantile-Based Intervals

Uncalibrated quantile regression undercovered on average (85.4%) and, more seriously, near failure (78.1%), where it overestimated the remaining life of 20.3% of engines, the worst safety record of all methods in distribution. Conformalising the quantiles restored average coverage (91.2%), but near-failure coverage remained unreliable, ranging from 80.0% on FD004 to 96.2% on FD003, and the correction widened intervals that were already wide: near failure, CQR intervals averaged 90.9 cycles, more than four times the Mondrian width. The quantile models produced their widest intervals for engines close to failure Figure 3, the opposite of what a maintenance planner needs.

4.5 Shift to a New Fault Mode

When intervals calibrated on FD001 were applied to FD003 engines, which can fail through an additional fault mode, all guarantees broke down Table 4 and Figure 5. Overall coverage fell to 68.4–75.1% for the tree ensembles and to about 50–60% for ridge regression and the MLP. Near failure, coverage collapsed to 36.9–50.0%, and the remaining life of 50–63% of near-failure engines was overestimated. Mondrian calibration did not help. Under the FD002 → FD004 shift, global split conformal prediction with tree ensembles retained near-failure coverage of 88–91%, but Mondrian calibration fell to 60–62%, because engines failing through the new mode were predicted to be healthier than they were and were assigned the narrow intervals of a healthy band. Across all shifted runs, 95.2% of miscoverage was overestimation of remaining life, compared with 63.7% in distribution: when the intervals failed under shift, they failed almost exclusively in the unsafe direction.

Table 4. Interval Performance under Fault-Mode Shift For the Tree-Ensemble Predictors (90% Nominal Coverage)

Shift	Predictor	Method	Coverage, All (%)	Coverage, RUL ≤ 30 (%)	Overestimated, All (%)	Overestimated, RUL ≤ 30 (%)	Width, RUL ≤ 30
FD001 → FD003	Gradient boosting	Split CP (all cycles)	68.4	36.9	29.9	63.1	43.6
		Mondrian CP	73.5	40.0	24.2	60.0	30.5
	Random forest	Split CP (all cycles)	70.5	41.9	27.9	58.1	44.0
		Mondrian CP	75.1	43.1	23.5	56.9	29.7
FD002 → FD004	Gradient boosting	Split CP (all cycles)	77.2	88.3	20.5	11.7	50.5
		Mondrian CP	75.4	60.0	22.8	38.9	25.3
	Random forest	Split CP (all cycles)	82.1	89.4	15.2	10.6	51.5
		Mondrian CP	79.3	62.3	19.5	37.7	26.7

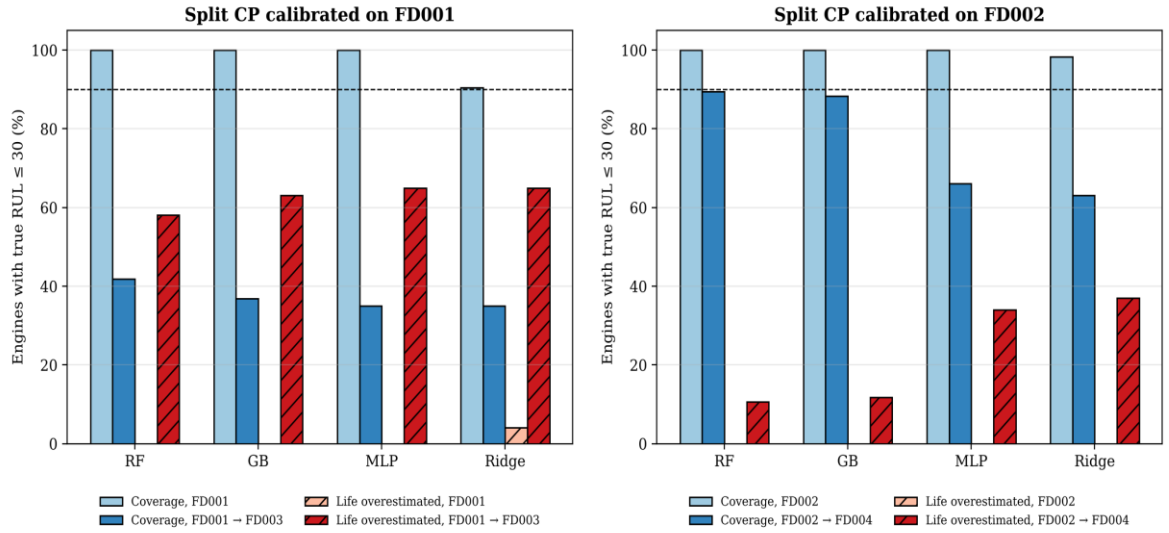


Figure 5. Near-Failure Coverage and Overestimation of Remaining Life for Global Split-Conformal Intervals Calibrated on Fd001 (Left) and Fd002 (Right), Evaluated in Distribution and on the Shifted Test Sets Fd003 and Fd004

4.6 Discussion

4.6.1 Principal Findings

Conformal prediction delivered what it promises, valid average coverage on exchangeable data, but this guarantee did not translate into intervals that are both safe and useful for the engines that matter most. Global intervals were uninformatively wide near failure and too narrow in the transition from healthy to degrading operation; band-wise calibration fixed the former for tree ensembles but depended on the quality of the underlying predictor; quantile-based intervals widened where they should narrow; and a change of fault mode removed the guarantee altogether, with errors that almost always overestimated the remaining life. These results extend the in-distribution findings of Javanmardi and Hüllermeier [17] and are consistent with the theoretical limits of marginal validity [15], [19] and exchangeability [20], [21].

4.6.2 Why Conformal Intervals Fail Near End of Life and under Shift

Two mechanisms explain the observations. First, the residuals of RUL predictors are strongly heteroscedastic: errors are small for healthy engines, whose target is capped, and for engines close to failure, whose degradation is unmistakable, but large in the transition, where the same sensor values are compatible with a wide range of remaining lives. A global quantile averages over these regimes, which simultaneously overcovers the ends and undercovers the middle. Calibrating within bands of predicted RUL recovers much of this structure, but because the band is assigned by the prediction, an engine whose degradation is underestimated receives a narrow, overconfident interval, which is exactly what happened with ridge regression and under shift. Second, a new fault mode changes how sensor signals evolve as an engine approaches failure. Models trained on a single fault mode see the unfamiliar pattern as less degraded than it is, so they predict too long a life, and the calibration set, drawn from the old fleet, offers no warning. This is why miscoverage under shift was overwhelmingly in the unsafe direction.

4.6.3 Implications for Safety-Critical Maintenance

Five practical recommendations follow. First, interval methods for prognostics should be evaluated not only by average coverage but by coverage, width and the rate of overestimation for engines close to failure, since that is where maintenance decisions are made. Second, band-wise (Mondrian) calibration is a simple and effective way to obtain informative near-failure intervals with accurate tree-ensemble predictors, but it should not be combined with a weak predictor. Third, uncalibrated quantile regression should not be used for safety decisions, and its conformalised version should be checked for width near failure. Fourth, a conformal guarantee should be considered void whenever the fleet differs from the calibration fleet in operating conditions or failure mechanisms; in such cases, recalibration on representative data, weighted conformal methods for covariate shift [20], [21] or explicit out-of-distribution monitoring are needed. Fifth, because errors under shift overestimate remaining life, asymmetric intervals or one-sided lower bounds that protect against late maintenance [18] deserve priority, in line with the asymmetric C-MAPSS scoring function [3].

4.6.4 Limitations

The study used a simulated benchmark; real engine fleets have additional sources of variability, such as maintenance actions and sensor faults. The RUL target was capped at 125 cycles, and coverage was evaluated against the capped true RUL, which is standard practice but makes intervals for healthy engines easier to satisfy. Calibration points from the same engine are not exchangeable with the single test point per engine, so the guarantee is only approximate even in distribution; the engine-subsampled design mitigates but does not remove this dependence. The point predictors used hand-crafted window features rather than deep sequence models, although their accuracy was comparable to published deep models. The number of near-failure test engines is small (20–61 per sub-dataset) and the number of calibration splits was limited to five or eight, so near-failure coverage estimates carry several percentage points of uncertainty. Finally, only two shift scenarios were studied, both involving an additional fault mode.

5. CONCLUSION

This study evaluated conformal and quantile-based prediction intervals for turbofan remaining useful life on all four C-MAPSS sub-datasets, with a focus on engines close to failure and on fault-mode shift. Global split conformal intervals met the nominal 90% coverage on average and covered almost all near-failure engines, but at a width of about 44 cycles, larger than the remaining life of those engines, and they undercovered engines in the transition to degradation. Calibrating within bands of predicted RUL halved near-failure widths for gradient boosting and random forests while keeping near-failure coverage at 96.5–97.8%, but failed with a less accurate linear predictor. Quantile-based intervals were widest exactly where narrow intervals are needed. Under a shift to a new fault mode, near-failure coverage fell to 37–50%, and 95% of all miscoverage overestimated the remaining life. Conformal prediction is a valuable tool for prognostics, but its marginal, exchangeability-based guarantee is not a safety guarantee; intervals intended for maintenance decisions should be assessed near end of life and under realistic shift, and combined with

predictors that are accurate in that regime. Future work will examine deep sequence models, asymmetric and one-sided conformal bounds, weighted conformal methods for operating-condition shift, and validation on real engine data.

Acknowledgments

The authors thank the NASA Ames Prognostics Center of Excellence for making the C-MAPSS data publicly available, the developers of scikit-learn and LightGBM, and the researchers whose work is cited in this article.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nuno Alexandre Soares Domingues	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Danish Ather		✓	✓		✓	✓	✓	✓		✓	✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable.

Data Availability

This study draws on the publicly available NASA C-MAPSS turbofan engine degradation data set. No new datasets were created. The analysis code is available from the corresponding author on reasonable request.

REFERENCES

[1] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation – A review on the statistical data driven approaches," European Journal of Operational Research, vol. 213, no. 1, pp. 1–14, 2011. doi.org/10.1016/j.ejor.2010.11.018

[2] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," Mechanical Systems and Signal Processing, vol. 104, pp. 799–834, 2018. doi.org/10.1016/j.ymssp.2017.11.016

[3] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in Proc. Int. Conf. Prognostics and Health Management, 2008, pp. 1–9. doi.org/10.1109/PHM.2008.4711414

[4] NASA Prognostics Center of Excellence, "Turbofan engine degradation simulation data set (C-MAPSS)," NASA Ames Research Center, PCoE Data Set Repository. www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/

[5] T. Wang, J. Yu, D. Siegel, and J. Lee, "A similarity-based prognostics approach for remaining useful life estimation of engineered systems," in Proc. Int. Conf. Prognostics and Health Management, 2008, pp. 1–6. doi.org/10.1109/PHM.2008.4711421

[6] F. O. Heimes, "Recurrent neural networks for remaining useful life estimation," in Proc. Int. Conf. Prognostics and Health Management, 2008, pp. 1–6. doi.org/10.1109/PHM.2008.4711422

[7] G. S. Babu, P. Zhao, and X.-L. Li, "Deep convolutional neural network based regression approach for estimation of remaining useful life," in Database Systems for Advanced Applications, Lecture Notes in Computer Science, vol. 9642. Cham: Springer, 2016, pp. 214–228. doi.org/10.1007/978-3-319-32025-0_14

[8] X. Li, Q. Ding, and J.-Q. Sun, "Remaining useful life estimation in prognostics using deep convolution neural networks," Reliability Engineering & System Safety, vol. 172, pp. 1–11, 2018. doi.org/10.1016/j.ress.2017.11.021

[9] S. Zheng, K. Ristovski, A. Farahat, and C. Gupta, "Long short-term memory network for remaining useful life estimation," in Proc. IEEE Int. Conf. Prognostics and Health Management, 2017, pp. 88–95. doi.org/10.1109/ICPHM.2017.7998311

[10] E. Ramasso and A. Saxena, "Performance benchmarking and analysis of prognostic methods for CMAPSS datasets," International Journal of Prognostics and Health Management, vol. 5, no. 2, 2014. doi.org/10.36001/ijphm.2014.v5i2.2236

[11] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019. doi.org/10.1038/s42256-019-0048-x

[12] M. Abdar et al., "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," Information Fusion, vol. 76, pp. 243–297, 2021. doi.org/10.1016/j.inffus.2021.05.008

[13] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York: Springer, 2005. doi.org/10.1007/b106715

[14] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in Machine Learning: ECML 2002, Lecture Notes in Computer Science, vol. 2430. Berlin: Springer, 2002, pp. 345–356. doi.org/10.1007/3-540-36755-1_29

[15] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1094–1111, 2018. doi.org/10.1080/01621459.2017.1307116

[16] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023. doi.org/10.1561/2200000101

[17] A. Javanmardi and E. Hüllermeier, "Conformal prediction intervals for remaining useful lifetime estimation," International Journal of Prognostics and Health Management, vol. 14, no. 2, 2023. doi.org/10.36001/ijphm.2023.v14i2.3417

[18] C. Robinson, "Remaining useful life estimation for aircraft engines with risk-aware prediction intervals via conformalized quantile regression," International Journal of Prognostics and Health Management, vol. 17, no. 1, 2026. doi.org/10.36001/ijphm.2026.v17i1.4724

[19] V. Vovk, "Conditional validity of inductive conformal predictors," in Proc. Asian Conf. Machine Learning, PMLR, vol. 25, 2012, pp. 475–490. proceedings.mlr.press/v25/vovk12.html

[20] R. J. Tibshirani, R. F. Barber, E. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in Advances in Neural Information Processing Systems 32, 2019, pp. 2530–2540. doi.org/10.48550/arXiv.1904.06019

[21] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," The Annals of Statistics, vol. 51, no. 2, pp. 816–845, 2023. doi.org/10.1214/23-AOS2276

[22] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd Int. Conf. Machine Learning, PMLR, vol. 48, 2016, pp. 1050–1059. proceedings.mlr.press/v48/gal16.html
[23] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems 30, 2017, pp. 6402–6413. doi.org/10.48550/arXiv.1612.01474
[24] Y. Romano, E. Patterson, and E. Candès, "Conformalized quantile regression," in Advances in Neural Information Processing Systems 32, 2019, pp. 3543–3553. doi.org/10.48550/arXiv.1905.03222
[25] R. Koenker and G. Bassett Jr., "Regression quantiles," Econometrica, vol. 46, no. 1, pp. 33–50, 1978. doi.org/10.2307/1913643
[26] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html
[27] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67, 1970. doi.org/10.1080/00401706.1970.10488634
[28] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi.org/10.1023/A:1010933404324
[29] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001. doi.org/10.1214/aos/1013203451
[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986. doi.org/10.1038/323533a0
[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations, 2015. doi.org/10.48550/arXiv.1412.6980
[32] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems 30, 2017, pp. 3146–3154. papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree
[33] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945. doi.org/10.2307/3001968

How to Cite: Nuno Alexandre Soares Domingues, Danish Ather (2026). Calibrated but not safe: Conformal prediction intervals for turbofan remaining useful life near end of life and under operating shift. Journal of Artificial Intelligence, Machine Learning and Neural Network, 6(2). 106-119. https://doi.org/10.55529/jaimlnn.62.106.1199

BIOGRAPHIES OF AUTHORS



Nuno Alexandre Soares Domingues, is an Adjunct Professor in the Department of Mechanical Engineering at Instituto Superior de Engenharia de Lisboa (ISEL), Instituto Politécnico de Lisboa, Lisbon, Portugal. His research addresses mechanical and energy engineering, including turbomachinery, energy systems and the condition and performance of engineering assets. He is interested in the maintenance and reliability of such assets and in how data-driven methods inform decisions about their remaining service life. Email: nuno.domingues@isel.pt



Danish Ather, is a Professor in the Department of Artificial Intelligence at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan. His research covers artificial intelligence methods and their application across engineering and health domains, including predictive modelling, uncertainty estimation and the evaluation of model explanations. He has a continuing interest in the methodology of AI research and in the standards by which such work is reviewed.