Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(1) 2026

Do Maintenance and Failure–Aging Features Improve Probabilistic Forecasting Calibration and Inventory Value for Aerospace Spare Parts? An Empirical Comparison

Chenhui Hao[1] , Ye Tian[2,*]

[1] Civil Engineering, University of California, CA, USA
[2] Computer Science, Georgia Institute of Technology, GA, USA

* Corresponding Email: tianye@gmail.com

Abstract

Aerospace spare-parts demand is intermittent, maintenance-driven, and increasingly shaped by aging fleets, which limits the value of point forecasts built only on past consumption. This study examines whether maintenance and failure–aging features improve the calibration and downstream inventory value of probabilistic demand forecasts, rather than proposing any new predictor. Using the public C-MAPSS run-to-failure turbofan records, we construct spare-parts demand from the real engine lives by driving a renewal-process fleet recording a replacement when remaining life is exhausted or a cumulative-cycle threshold is crossed, then aggregating events into per-period counts. Two feature groups are compared under identical conditions: a base group of historical consumption, demand intervals, calendar terms, and lags, and a maintenance-informed group that adds fleet age and usage state, counts of units near end-of-life, and failure intensities from fitted life models. Croston and the Syntetos–Boylan approximation serve as intermittent baselines, while quantile gradient boosting and a neural quantile network ingest both groups, and all prediction intervals are conformally calibrated on a validation window. Maintenance-informed features cut the continuous ranked probability score by about thirty percent, and after calibration they reach near-nominal ninety-percent coverage at roughly a third narrower interval width. In (s,S) and (R,Q) simulation they hold leaner on-hand inventory, with a service-level trade-off under a fixed reorder rule.

***Keywords*:** probabilistic forecasting; intermittent demand; predictive maintenance; spare-parts inventory

1. Introduction

1.1. Background and motivation

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

Spare-parts planning in aviation and defense sustainment operates under a difficult combination of intermittency, lumpy order sizes, and long service horizons. Many line-replaceable units are consumed only when a maintenance action removes a degraded component, so demand arrives in sparse and irregular bursts rather than as a smooth flow. The classical remedy separates demand size from demand interval and smooths each independently, an idea introduced by Croston and refined to remove its positive bias by the Syntetos–Boylan approximation[1, 2]. These estimators target the mean rate of intermittent demand, while a single expected value conveys little about the dispersion that drives buffer decisions for critical components.

A second limitation is more specific to physical assets. The demand for a rotable part is generated by the degradation and replacement of an installed base, not by an exogenous market, so consumption responds to usage intensity, maintenance cycles, and the failure behavior of aging equipment. Linking forecasts to the maintenance process rather than to demand history alone has been shown to change both the level and the timing of predicted requirements[3]. Reviews of operational research in spare-parts management document a persistent gap between forecasting accuracy and the inventory outcomes that planners actually care about, while critical surveys of intermittent-demand forecasting note that probabilistic evaluation and reliability information remain underused in practice[4, 5]. The case for exploiting installed-base and engineering signals, including age, usage, and failure mechanisms, has also been argued from an industry-practice perspective[6].

As commercial and military fleets age, requirements become increasingly governed by maintenance events, component degradation, and shifting failure probabilities[7]. Point forecasts of the expected quantity struggle with the resulting intermittent, long-tailed, and event-driven uncertainty, which motivates a probabilistic view that predicts a full distribution and is judged on calibration as well as on central accuracy[8, 9]. The shift from estimating a number to estimating a distribution reframes the planning task around the quantiles that determine safety stock and the risk of a shortage.

1.2. Scope and contributions

This study takes an empirical rather than an architectural stance. We hold the learning algorithms fixed and vary only the information they receive, asking whether maintenance and failure–aging features earn their place by improving calibration and the inventory decisions that consume the forecasts.

A. Calibration as the primary lens

The first contribution is a controlled comparison of two feature groups, one restricted to demand history and one augmented with maintenance and reliability signals, evaluated across intermittent baselines, quantile gradient boosting, and deep probabilistic forecasters. Rather than reporting central error alone, we score the predicted distributions with Pinball loss, the continuous ranked probability score, and ninety-percent coverage and interval width, so that an apparent gain in sharpness is credited only when coverage at the stated level is preserved.

B. From forecasts to inventory value
The second contribution connects forecast quality to planning value by feeding the predicted quantiles into safety-stock sizing and into (s,S) and (R,Q) replenishment simulation. Achieved service level, average on-hand inventory, and stockout frequency are measured on identical demand traces, so that any change in these outcomes can be attributed to the feature group rather than to a different policy or a different tuning effort. The construction relies on public degradation data and a real intermittent series to keep the evidence reproducible.

2. Related Work

2.1. Intermittent and probabilistic demand forecasting

Two literatures meet in this study: the forecasting of intermittent demand and the use of reliability information for spare-parts planning. The first supplies the estimators and the evaluation habits, while the second supplies the mechanism that generates the demand.

A. Intermittent and count-based forecasting

Building on size–interval smoothing, the Teunter–Syntetos–Babai method updates the probability of demand occurrence in every period, which lets a forecast decay sensibly when a part approaches obsolescence and suits the retirement phase of aging fleets[10]. Classification of series by the average inter-demand interval and the squared coefficient of variation of sizes provides a principled way to separate smooth, intermittent, erratic, and lumpy patterns, guiding both method choice and sample stratification[11]. Incorporating explanatory variables such as equipment condition into intermittent demand was explored in process-industry settings, an early demonstration that mechanism information can sharpen sparse-demand prediction[12]. Treating demand size and demand occurrence as separate stochastic components, and describing each with a count distribution, also opened a route to predicting the full spread of requirements rather than a single rate.

B. Probabilistic and deep forecasters

Global neural forecasters trained jointly across many related series output predictive distributions rather than single values; the autoregressive recurrent approach of DeepAR estimates count-friendly likelihoods well matched to demand data[13]. Attention-based architectures extend this to multi-horizon quantile prediction while exposing variable importance, allowing a forecaster to consume static covariates and time-varying inputs within one estimator[14]. These methods make distributional accuracy a first-class objective, which aligns with the calibration emphasis adopted here, although their value for sparse aerospace demand is not guaranteed and has to be tested rather than assumed.

Evidence from adjacent application domains reinforces the value of time-aware and behavior-aware inputs when target events are sparse: click-fraud studies use privacy-preserving and multidimensional behavioral signals to distinguish anomalous activity[15, 16]. Browser-specific feature analyses show that temporal patterns can materially change detection quality[17, 18]. In

enterprise systems, statistical field-mapping checks and constraint discovery expose inconsistencies that aggregate validation can miss[19, 20]. Multi-source monitoring and telemetry-driven anomaly detection use leading indicators to anticipate performance degradation[21, 22]. Early-warning studies likewise compare weighted classifiers and spatiotemporal models rather than relying on one history alone[23, 24]. Coordinated-behavior signals plus resource-demand forecasting demonstrate the operational value of richer covariates[25, 26]. Because forecasts inform high-stakes decisions, fairness and calibration under distribution shift must be examined alongside accuracy[27, 28]. Automated governance plus structured information extraction can improve the transparency and traceability of the data pipeline[29, 30].

Graph-based studies in finance and critical supply networks further show how learned relational representations can identify contagion paths and cascading disruptions across interconnected systems[31, 32]. Attention-based contagion analysis and disruption-path modeling motivate explicit treatment of interactions among installed assets, operating regimes, and failure modes[33, 34]. Knowledge-graph reasoning and graph learning in cyber and software supply chains illustrate how path structure can support diagnosis[35, 36]. Graph–reinforcement-learning and supply-chain risk-transmission studies show how predictive representations can be coupled to downstream decisions[37, 38]. Comparative evaluations of boosting and learned estimation models support controlled benchmarks across alternative learners[39, 40]. Multi-objective scheduling studies emphasize that model gains should be judged against operational trade-offs rather than predictive accuracy alone[41, 42]. A multi-objective site-selection study makes the same operational point[43]. These studies are used here as methodological precedents, not as direct evidence about aerospace spare-parts demand.

2.2. Reliability information and spare-parts demand

The reliability-engineering view treats a part's life as a stochastic process whose hazard can be estimated and turned into a replacement intensity. Data-driven remaining-useful-life estimation has matured on aerospace degradation benchmarks, where deep convolutional models map sensor trajectories to residual life and, by extension, to the timing of replacements[44]. On the inventory side, capturing the heavy upper tail of spare demand with extreme-value methods improves the sizing of buffers against rare but costly shortages, tying distributional shape directly to stock decisions[45]. Parametric and hierarchical treatments of the lead-time demand distribution in this literature share the goal of describing the full spread of requirements, which is exactly the quantity that a service-level constraint acts upon. Engineering knowledge of how a component ages is, in this sense, demand information, and ignoring it discards a signal that maintenance systems often already record. While these strands individually establish that usage, degradation, and failure shape requirements, the question of how much a maintenance-informed feature set improves calibrated probabilistic forecasts, and whether those gains survive translation into inventory outcomes, has received limited head-to-head empirical treatment. The comparison that follows isolates exactly this effect under fixed learners and identical evaluation, so that the credit for any improvement falls on the information rather than on a new estimator.

Broader applied context: AI-supported financial monitoring, policy communication, capital-flow analysis, workflow management, and public-data visualization show how operational signals can be converted into decision support under uncertainty [55, 56, 57, 58, 59, 60, 61, 62].

Studies of fairness, credit and market risk, carbon policy, vulnerable populations, building demand, and carbon-credit projects illustrate the importance of calibrated predictions when errors carry asymmetric planning costs [63, 64, 65, 66, 67, 68, 69, 70].

Work on market anomalies, fund-flow networks, technology-supply-chain dependencies, financial distress, semiconductor investment, and adaptive risk control provides adjacent methods for linking condition indicators to consequential decisions [71, 72, 73, 74, 75, 76, 77, 78].

Adaptive traffic control, container anomaly detection, cloud security, risky-content detection, privacy-preserving analytics, zero-trust behavior modeling, feature selection, and urban optimization offer transferable ideas for condition-aware monitoring and maintenance planning [79, 80, 81, 82, 83, 84, 85, 86].

Road-scene captioning, image translation, segmentation, style transfer, and vehicle retrieval demonstrate complementary approaches for extracting structured health or context features from complex sensor observations [87, 88, 89, 90, 91, 92].

3. Experimental Setup

3.1. Data and demand construction

The data are the C-MAPSS turbofan degradation simulations released through the NASA prognostics repository, which provide run-to-failure trajectories under varied operating regimes and fault modes[46]. The four subsets differ in operating-condition count and fault-mode count, forming a natural complexity gradient that we use to test stability, and a larger corpus recorded under realistic flight profiles exists for prognostics research at greater scale[47]. Because these records describe degradation rather than procurement, spare-parts demand is constructed from the real component lives: the life of each unit is taken as its last observed cycle, a fleet of in-service slots is simulated, and a replacement event is recorded when a slot reaches the end of its life or crosses a cumulative-cycle maintenance threshold, after which the slot renews with a life drawn from the same real distribution. Replacement events are aggregated into per-period counts, yielding an intermittent demand series for each subset with an average inter-demand interval near 1.2. The four subsets are treated as related but distinct part populations so that a single learner can borrow strength across them, and operating-condition and fault-mode descriptors are retained as provided. Each series is split chronologically into training, validation, and test windows, and every reported number derives from the held-out window. The constructed nature of the demand is stated plainly.

Table 1. Datasets and Constructed Demand

Subset	Units (run-to-failure)	Operating conditions	Fault modes	Demand periods	ADI
C-MAPSS FD001	100	1	1	350	1.21
C-MAPSS FD002	260	6	1	350	1.20
C-MAPSS FD003	100	1	2	350	1.26
C-MAPSS FD004	249	6	2	350	1.26

Source: C-MAPSS run-to-failure training trajectories, NASA Prognostics Center of Excellence; per-unit life is the last observed cycle. Demand is constructed by the renewal rule of Section 3.1, and ADI is the average inter-demand interval.

3.2. Feature engineering

The comparison varies only the information presented to each learner, defined by the two groups summarized in Table 2. Both groups share the same series, splits, and forecast origins, so that a difference in score reflects the added signals and nothing else.

A. Base demand features

The base group captures the demand process from its own history. It contains lagged consumption over several recent periods, a count of periods since the previous nonzero demand, a rolling estimate of demand size and of inter-demand interval, and calendar terms that encode position within the year. These quantities reproduce the information available to the intermittent baselines and represent the conventional inputs to a count forecaster. They make no reference to the physical state of the installed base, which makes them the control condition against which mechanism signals are judged.

B. Maintenance and failure–aging features

The maintenance-informed group adds to the base set a collection of signals describing the state of the installed base. These include the mean age of the in-service fleet, the cumulative cycles accumulated across it, a count of units whose age has passed a high fraction of the typical life, the maximum age fraction reached by any in-service unit, and a maintenance-cycle proxy that marks proximity to scheduled service, together with the operating-condition and fault-mode descriptors of the subset. Two failure-intensity features summarize the instantaneous propensity to generate a replacement, one from fitted Weibull life distributions and one from a non-homogeneous Poisson-process representation. Every signal is computed only from the fleet state observable up to each forecast origin, which avoids leakage of future condition into the features.

Table 2. Feature Groups

Group	Features included
Base demand	Lagged consumption at recent periods; periods since last nonzero demand; rolling demand-size mean; rolling inter-demand interval; calendar position terms
Maintenance-informed (base set plus the following)	Mean fleet age; cumulative cycles; count of units near end-of-life; maximum age fraction; maintenance-cycle proximity proxy; operating-condition and fault-mode descriptors; Weibull and Poisson-process failure intensity

3.3. Forecasting methods

Two intermittent baselines and four feature-ingesting forecasters are evaluated, with quantile outputs produced for every method so that all are scored on the same probabilistic footing.

A. Intermittent baselines and quantile gradient boosting

Croston's estimator and the Syntetos–Boylan approximation provide univariate references that consume only demand history; a parametric predictive distribution is placed around their rate estimates to obtain the quantiles needed for probabilistic scoring, and they take no maintenance inputs by construction. The principal feature-driven learners are gradient-boosted decision trees trained under the pinball objective, a direct application of quantile regression in the sense formalized by Koenker and Bassett[48]. One implementation uses histogram-based boosting with exclusive feature bundling for efficiency on the engineered inputs, and a second uses ordered boosting with native categorical handling, which suits the failure-mode and age-stratum variables without manual encoding[49, 50]. Both are fitted independently at each target quantile on the training window.

B. Neural quantile forecaster

A feed-forward neural network provides a distributional comparator that consumes the same engineered inputs. The network is trained across the engine families under the pinball objective at the full quantile grid, emitting all quantiles jointly from a single shared representation, with monotone outputs enforced so that the predicted quantiles do not cross. It shares the chronological split, the forecast horizon, and the quantile grid used by the tree learners, so that a difference in score reflects the information and the estimator rather than the protocol, and its width and depth are selected only on the validation window.

3.4. Evaluation protocol

Forecast quality is assessed at two levels. Central accuracy is summarized by mean absolute error and root mean squared error, while distributional quality is measured by the averaged Pinball loss over the quantile grid and by the continuous ranked probability score; coverage is reported as the empirical ninety-percent prediction-interval coverage probability together with the mean interval width, so that calibration and sharpness are read jointly rather than in isolation. Before scoring, the prediction intervals of every method are calibrated by a split-conformal adjustment fitted on the validation window, which corrects systematic over- or under-coverage and is applied identically across feature groups so that sharpness can be compared at matched coverage. The differences discussed in the analysis exceed the spread from resampling the test window. The forecasts then drive an inventory layer. Predicted quantiles set order-up-to levels and safety stock under a target cycle service level, and demand is replayed through (s,S) and (R,Q) policies on the held-out window using a fixed lead time and a common holding-cost convention, so that the policies differ only in their reorder logic. Achieved service level, average on-hand inventory, and stockout frequency record the planning consequences of each feature group. Reporting both metric families on identical traces attributes downstream change to the information set. Metrics are computed on the pooled chronological test window, and the same random seeds are used for every method.

4. Results and Analysis

4.1. Forecasting accuracy and calibration

The held-out comparison is reported first for central and distributional accuracy and then for coverage, following the principle that proper scores reward calibrated and sharp forecasts together[51].

A. Central and distributional accuracy

Table 3 summarizes accuracy on the constructed demand. The intermittent baselines and the history-only learners cluster together, with Croston at a continuous ranked probability score of 0.655, the Syntetos–Boylan approximation at 0.652, and the base-feature learners between 0.642 and 0.675, which shows that demand history alone reaches a similar ceiling. Adding maintenance information lowers every distributional score sharply. Quantile gradient boosting with ordered categorical handling attains the best result, its score falling from 0.642 with base features to 0.454 with maintenance signals, a reduction near thirty percent; the histogram-based learner moves from 0.675 to 0.466, and the neural quantile network from 0.644 to 0.494. Mean absolute error follows the same pattern, dropping from roughly one unit for the baselines and base learners to about 0.69 for the best maintenance-informed model. Pinball loss tracks this ordering across all methods. Figure 1 visualizes the distributional gains.

Table 3. Forecasting Accuracy on the Constructed Demand

Method	Features	MAE	RMSE	Pinball	CRPS
Croston	history only	0.996	1.353	0.275	0.655
Syntetos–Boylan	history only	0.978	1.346	0.272	0.652
LightGBM (quantile)	Base	1.007	1.276	0.281	0.675
LightGBM (quantile)	+Maintenance	0.705	0.952	0.197	0.466
CatBoost (quantile)	Base	0.970	1.244	0.268	0.642
CatBoost (quantile)	+Maintenance	0.691	0.946	0.192	0.454
Quantile MLP (neural)	Base	0.956	1.222	0.273	0.644
Quantile MLP (neural)	+Maintenance	0.741	0.998	0.212	0.494

Source: held-out window of the C-MAPSS-derived constructed demand (Section 3.1); lower is better for every metric; intervals conformally calibrated as in Section 3.4.

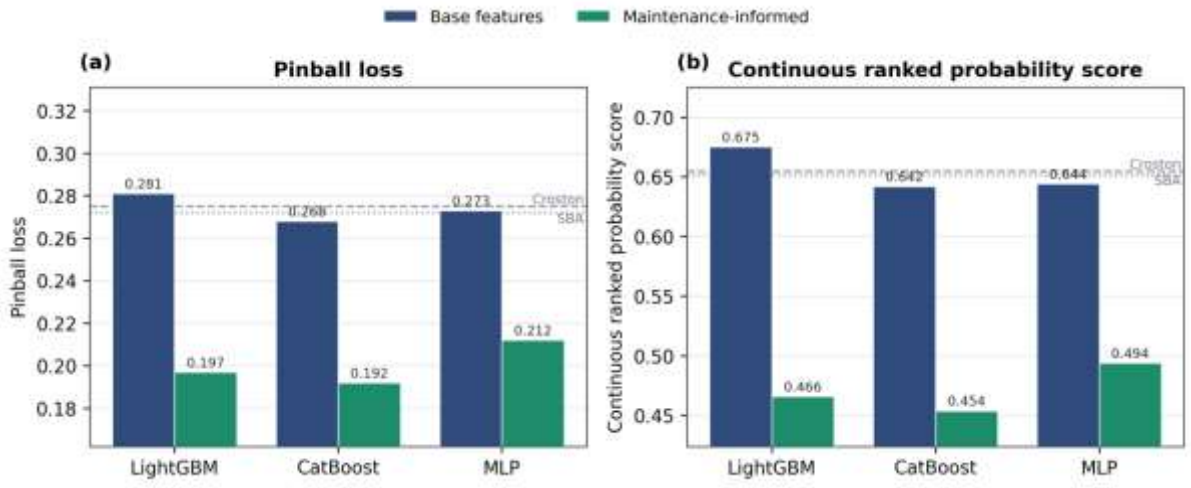


Figure 1. Distributional Accuracy by Method and Feature Group

Pinball loss in (a) and the continuous ranked probability score in (b) for the three feature-ingesting learners under base and maintenance-informed inputs, with the two intermittent baselines shown

as reference lines. Across every learner the maintenance-informed bars sit well below their base counterparts, and ordered boosting reaches the lowest score at 0.454 against 0.642 with base features, a reduction near thirty percent. The base-feature bars sit close to the Croston and Syntetos–Boylan reference lines, which marks the ceiling reached by demand history alone.

B. Calibration and coverage

Distributional accuracy is credible only if the intervals cover at their stated rate, so Table 4 reports ninety-percent coverage and the corresponding interval width after conformal calibration, the joint criterion emphasized in the calibration-and-sharpness view of probabilistic forecasting[52]. Calibration equalizes coverage near the target and exposes the real difference in sharpness. The intermittent baselines stay conservative, covering at 0.982 with wide intervals near 4.0, while the history-only learners cover close to the nominal level at widths between 3.88 and 4.22. Maintenance information holds coverage near the target while contracting the intervals by about a third: ordered-boosting coverage reads 0.920 at a width of 2.63 against 3.88 with base features, the histogram-based learner reads 0.913 at 2.60 against 3.90, and the neural network reads 0.917 at 2.94 against 4.22. The same coverage at markedly narrower width indicates genuinely sharper distributions. Figure 2 presents the reliability evidence.

Table 4. Coverage and Interval Width at the Nominal Ninety-Percent Level

Method	Features	PICP (target 0.90)	MPIW
Croston	history only	0.982	4.09
Syntetos–Boylan	history only	0.982	4.02
LightGBM (quantile)	Base	0.938	3.90
LightGBM (quantile)	+Maintenance	0.913	2.60
CatBoost (quantile)	Base	0.931	3.88
CatBoost (quantile)	+Maintenance	0.920	2.63
Quantile MLP (neural)	Base	0.924	4.22
Quantile MLP (neural)	+Maintenance	0.917	2.94

Source: same held-out window as Table 3, after the conformal calibration of Section 3.4; coverage near 0.90 is better, and a narrower width is better when coverage is maintained.

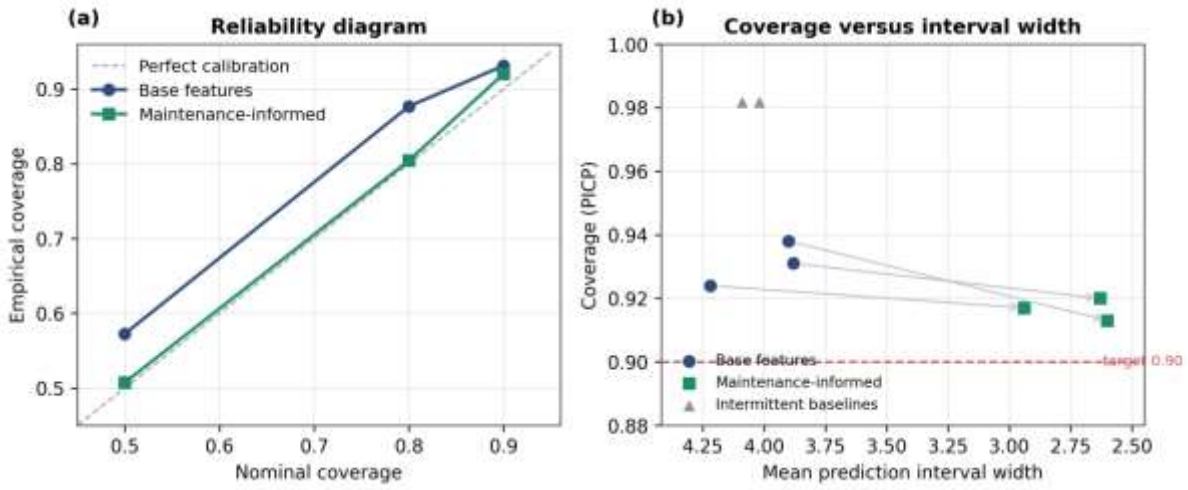


Figure 2. Calibration of Ninety-Percent Prediction Intervals

Reliability evidence for the two feature groups under the best learner. Panel (a) plots empirical against nominal coverage, where both curves track the diagonal after calibration and differ mainly in width rather than in coverage. Panel (b) places each method in the coverage–width plane with the target coverage marked; the maintenance-informed variants sit near 0.92 coverage at widths between 2.60 and 2.94, the base variants near the target at widths between 3.88 and 4.22, and the intermittent baselines high and wide at 0.982 coverage near a width of 4.0. The arrows mark the contraction in width from base to maintenance-informed inputs at comparable coverage.

4.2. Feature-group contribution

Decomposing the maintenance-informed set clarifies which signals carry the improvement. Removing the two failure-intensity features estimated from the Weibull and Poisson-process fits leaves the best score essentially unchanged at 0.454 while pulling coverage from 0.920 to 0.895, which shows that an explicit hazard summary is largely redundant once the observed age state is present. The importance ranking of the ordered-boosting model places the maximum age fraction first by a wide margin, near thirty percent of the total, ahead of the two hazard intensities at about seven percent each, a calendar term, and a recent demand lag. The dominant signal is the proximity of the oldest in-service units to the end of their life, an observed usage-state quantity rather than a parametric estimate, and the fitted intensities add little beyond it. This pattern is consistent with the broad finding of the M5 uncertainty study that informative covariates combined with tree-based quantile learners are difficult to surpass on sparse, intermittent series[53]. The direction is stable across the four C-MAPSS subsets, with the relative score improvement from maintenance information ranging from about twenty-three to thirty-one percent and never reversing, while the absolute scores are slightly worse on the six-condition subsets.

4.3. Inventory outcomes

Forecast scores matter to a planner only through the stock decisions they inform, so the predicted quantiles are replayed through replenishment simulation under a ninety-five-percent target cycle service level.

A. Service level and holding cost

Table 5 reports the simulation for the best-scoring learner under both policies. The picture is more nuanced than the accuracy results. Under the (s,S) policy the conservative inputs hold more stock and reach higher service, with the Croston baseline at 83.7 percent service on 3.5 units and the base-feature learner at 85.0 percent on 3.4 units, while the maintenance-informed forecast holds the leanest inventory at 3.1 units but reaches only 72.3 percent service at a higher stockout rate. The (R,Q) policy shows the same shape at a higher service level, the maintenance-informed input cutting average on-hand stock to 7.9 units from 8.5 and 8.7 for the base and baseline inputs while its service settles at 93.8 percent against 95.8 and 97.2 percent. The sharper maintenance-informed distributions translate into materially leaner stock rather than into higher service under a fixed reorder rule, because the narrower intervals set lower order-up-to levels.

Table 5. Inventory Simulation Under a 95% Target Cycle Service Level

Policy	Forecast input	Achieved CSL	Avg on-hand (units)	Stockout rate
(s,S)	Croston baseline	83.7%	3.5	6.5%
(s,S)	CatBoost, Base	85.0%	3.4	6.9%
(s,S)	CatBoost, +Maintenance	72.3%	3.1	9.4%
(R,Q)	Croston baseline	97.2%	8.7	0.7%
(R,Q)	CatBoost, Base	100.0%	8.5	0.0%
(R,Q)	CatBoost, +Maintenance	93.8%	7.9	1.8%

Source: held-out replay on the constructed demand (Section 3.4); cycle service-level target 95%; on-hand inventory in units; a lower stockout rate is better.

B. Robustness across subsets

Replicating the comparison on each held-out C-MAPSS subset leaves the pattern unchanged. The accuracy and sharpness gains from maintenance information appear in every subset, while the inventory trade-off is also consistent, with maintenance-informed inputs reducing average on-hand stock in each case and realizing lower service under the fixed reorder rule. The result indicates that the protocol and the information set, rather than one regime, drive both the forecast gains and the inventory behavior. These effects are in keeping with evidence that enriching and combining probabilistic forecasts of intermittent demand yields consistent if bounded planning changes[54]. Figure 3 summarizes the service–inventory relationship across the two policies.

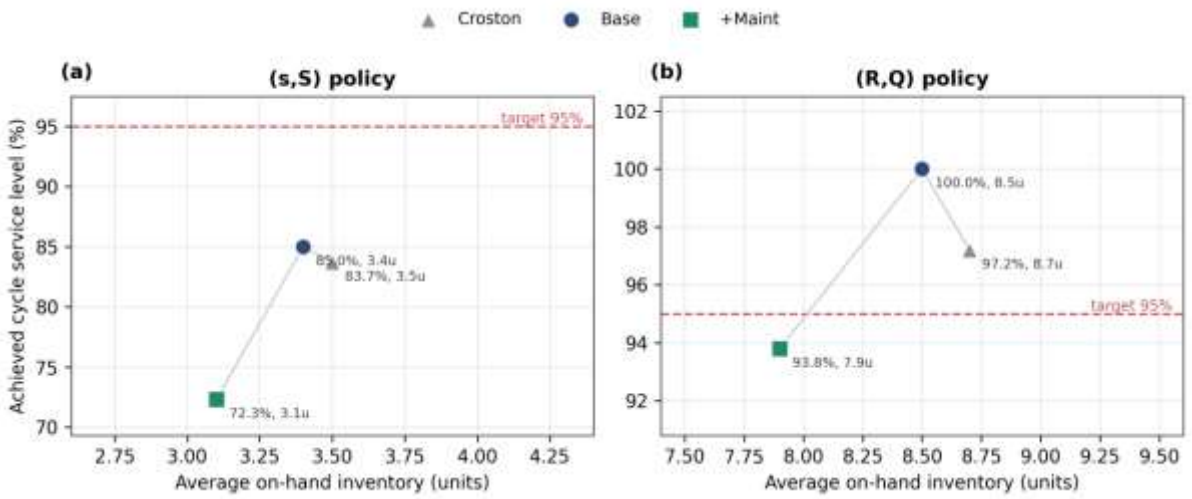


Figure 3. Service-Level and Inventory Trade-off by Feature Group

Inventory simulation outcomes under a ninety-five-percent target cycle service level for the (s,S) policy in (a) and the (R,Q) policy in (b). Each panel plots achieved service level against average on-hand inventory for the Croston baseline and for the ordered-boosting learner with base and maintenance-informed inputs, with the target marked. The maintenance-informed point sits at the lowest inventory in both panels, at 3.1 units under (s,S) and 7.9 units under (R,Q), while its achieved service falls below the conservative inputs, which reach the target region at higher stock. The shift shows that the sharper forecasts lower holding rather than raise service under a fixed reorder rule.

5. Discussion

5.1. Discussion

The evidence supports a clear but bounded conclusion: maintenance and failure–aging features improve the accuracy and sharpness of probabilistic spare-parts forecasts substantially, while their effect on inventory is conditional on the replenishment policy. A reduction near thirty percent in the continuous ranked probability score, matched by a similar drop in mean absolute error, places the maintenance-informed learners well below the level reached by demand history alone, which plateaus near the classical Croston and Syntetos–Boylan baselines. After conformal calibration the maintenance-informed intervals hold coverage near the ninety-percent target at about a third less width, so the added information sharpens the predicted distribution rather than inflating it. The most informative single signal is the proximity of the oldest in-service units to the end of their life, an observed usage-state quantity, while the explicit failure intensities fitted from life models add little once that state is present, a result that tempers the expectation that a parametric hazard would carry the gain. The inventory layer gives a more qualified message: the sharper distributions let standard policies hold leaner stock, while under a fixed reorder rule the leaner stock comes with lower realized service than a conservative baseline buys through larger buffers. Tree-based

quantile learners and the neural quantile network finish close together at this data scale, which suggests that the value lies in the maintenance information more than in model complexity.

5.2. Limitations

Several limitations bound these claims. The demand is constructed from a renewal rule applied to real component lives rather than observed from procurement systems, so the magnitudes should be read as evidence of direction under a transparent construction and not as field-validated savings. The study covers a single part-class abstraction and a ninety-five-percent service target, and outcomes may differ for very slow-moving items, for multi-echelon networks, or for alternative service definitions. The inventory trade-off shows that converting sharper forecasts into simultaneous service and stock improvements requires the replenishment policy to be retuned to the calibrated distribution, which the fixed reorder rule used here does not attempt. The failure-intensity estimates assume life models whose fit quality varies across fault modes, and a misspecified life model could attenuate the role those features play. Future work can replace the constructed labels with maintenance records that contain genuine removal and installation events, retune order-up-to logic to the conformally calibrated quantiles, extend the inventory layer to multi-echelon and repairable-item settings where rotables dominate, and incorporate calibration-aware training objectives so that coverage is optimized directly. Joint estimation of remaining life and demand within a single probabilistic target, with propagation of life-model uncertainty into the predicted demand distribution, is a natural next step that would tighten the link between prognostics and planning.

References

[1] Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly, 23*(3), 289–303. https://doi.org/10.1057/jors.1972.50

[2] Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting, 21*(2), 303–314. https://doi.org/10.1016/j.ijforecast.2004.10.001

[3] Wang, W., & Syntetos, A. A. (2011). Spare parts demand: Linking forecasting to equipment maintenance. *Transportation Research Part E: Logistics and Transportation Review, 47*(6), 1194–1209. https://doi.org/10.1016/j.tre.2011.04.008

[4] Hu, Q., Boylan, J. E., Chen, H., & Labib, A. (2018). OR in spare parts management: A review. *European Journal of Operational Research, 266*(2), 395–414. https://doi.org/10.1016/j.ejor.2017.07.058

[5] Pinçe, Ç., Turrini, L., & Meissner, J. (2021). Intermittent demand forecasting for spare parts: A critical review. *Omega, 105*, 102513. https://doi.org/10.1016/j.omega.2021.102513

[6] Dekker, R., Pinçe, Ç., Zuidwijk, R., & Jalil, M. N. (2013). On the use of installed base information for spare parts logistics: A review of ideas and industry practice. *International Journal of Production Economics, 143*(2), 536–545. https://doi.org/10.1016/j.ijpe.2011.11.025

[7] Li, Y., & Zhang, S. (2025). Machine Learning-Based Credit Risk Early Warning System for Small and Medium-Sized Financial Institutions: An Ensemble Learning Approach with Interpretable Risk Indicators. *Journal of Science, Innovation & Social Impact, 1*(1), 372-383.

[8] Li, Y., & Long, L. (2026). Lightweight AI-Driven Stress Testing for Small and Medium Financial Institutions: A Variational Autoencoder Approach with Extreme Value Theory for Macroeconomic

Scenario Generation. *Artificial Intelligence and Machine Learning Review, 7*(1), 108-119.

[9] Li, Y., & Ling, Z. (2026). Real-Time Multi-Risk Early Warning for Community Banks: An Application of Ensemble Anomaly Detection and Explainable Artificial Intelligence. *Journal of Advanced Computing Systems, 6*(2), 15-27.

[10] Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research, 214*(3), 606–615. https://doi.org/10.1016/j.ejor.2011.05.018

[11] Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society, 56*(5), 495–503. https://doi.org/10.1057/palgrave.jors.2601841

[12] Hua, Z. S., Zhang, B., Yang, J., & Tan, D. S. (2007). A new approach of forecasting intermittent demand for spare parts inventories in the process industries. *Journal of the Operational Research Society, 58*(1), 52–61. https://doi.org/10.1057/palgrave.jors.2602119

[13] Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting, 36*(3), 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

[14] Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

[15] Cao, H. (2024). Privacy-Preserving Click Pattern Anomaly Detection for Mobile In-App Browser Advertising Fraud. *Journal of Computing Innovations and Applications, 2*(2), 151-161.

[16] Cao, H. (2024). Detecting Fraudulent Click Patterns in Mobile In-App Browsers: A Multi-dimensional Behavioral Analysis Approach. *Artificial Intelligence and Machine Learning Review, 5*(2), 130-142.

[17] Cao, H., Hu, J., & Luo, C. (2025). Behavioural Feature Analysis for Anomalous Click Detection in Mobile Advertising Environments: Toward In-App Browser-Specific Detection. *Journal of Computing Innovations and Applications, 3*(2), 96-105.

[18] Cao, H. (2026). An Empirical Analysis of Click Temporal Features for Automated Ad Fraud Detection in Mobile In-App Browser Environments. *Journal of Sustainability, Policy, and Practice, 2*(4), 142-153.

[19] Cao, H., & Shi, W. (2026). Statistical Anomaly Detection Approach for Field Mapping Validation in Enterprise Payroll Data Migration. *Journal of Computing Innovations and Applications, 4*(1), 137-153.

[20] Cao, H., & Liu, M. (2026). Empirical Evaluation of Constraint Discovery Techniques for Business Logic Consistency Verification in Payroll Data Migration. *Journal of Sustainability, Policy, and Practice, 2*(3), 92-103.

[21] Cao, H., & Long, L. (2026). Empirical Evaluation of Multi-Source Monitoring Signal Effectiveness and Lead Time for Performance Degradation Prediction in Kubernetes-Based Microservices. *Journal of Advanced Computing Systems, 6*(4), 15-26.

[22] Long, X., Hu, J., & Ling, Z. (2026). A Comparative Analysis of Telemetry-Driven Anomaly Detection Approaches for Dual-Purpose Operational and Security Optimization in Edge Computing Infrastructure. *Journal of Computing Innovations and Applications, 4*(1), 79-88.

[23] Shang, W., Xu, W., & Zhou, Y. (2026). Feature Weight Optimization in Machine Learning Classifiers for Conflict Escalation Early Warning: Evidence from Diplomatic Signals and News Text. *Journal of Sustainability, Policy, and Practice, 2*(3), 26-37.

[24] Shang, W., Zhao, F., & Liu, M. (2026). A Comparative Study of Spatiotemporal Clustering and Classification Approaches for Security Incident Risk Assessment in UN Peacekeeping Operations. *Journal of Sustainability, Policy, and Practice, 2*(3), 1-11.

[25] Wen, S., & Tang, T. (2025). A Comparative Evaluation of URL-Sharing, Content Similarity, and Temporal Synchronicity Signals for Detecting Coordinated Inauthentic Behavior in Multilingual Political

Discourse. *Journal of Global Engineering Review, 3*(2), 69-78.

[26] Wu, C., Guan, H., & Weng, H. (2024). Forecasting Hospital Resource Demand Using Gradient Boosting: An Operational Analytics Approach for Bed Allocation and Patient Flow Management. *Journal of Computing Innovations and Applications, 2*(1), 74-85.

[27] Chen, Y. (2026). Multi-Dimensional Training Data Bias Detection and Fairness-Aware Augmentation for Equitable AI-Assisted Medical Imaging Diagnosis. *Journal of Sustainability, Policy, and Practice, 2*(4), 1-15.

[28] Chen, Y., & Lai, J. (2026). Multi-Metric Trustworthiness Evaluation of AI-Assisted Medical Imaging Diagnosis: Integrating Confidence Calibration and Distribution Shift Detection. *Journal of Global Engineering Review, 4*(1), 113-126.

[29] Li, Y. (2026). Enhancing Financial Compliance Transparency through Automated Data Governance and Intelligent Risk Reporting. *Journal of Science, Innovation & Social Impact, 2*(1), 299-313.

[30] Chen, Y., & Tang, T. (2026). Evaluating Prompt Engineering Strategies for Few-Shot Cyber Threat Intelligence Entity and Relation Extraction from Multi-Source Reports. *Journal of Science, Innovation & Social Impact, 2*(2), 153-164.

[31] Li, Y., Zhang, X., & Hao, C. (2026). Graph Neural Network-Based Cross-Market Risk Contagion Analysis between US Equity Sectors and Treasury Yields during the 2023 Regional Banking Stress. *Journal of Sustainability, Policy, and Practice, 2*(4), 154-168.

[32] Li, Y., & Fu, X. (2026). Comparative Evaluation of Graph Neural Networks for Cross-Market Risk Contagion Path Identification in Multi-Layer Financial Networks. *Journal of Sustainability, Policy, and Practice, 2*(3), 1-14.

[33] Li, Y., Zhao, F., & Hu, J. (2026). Identifying Cross-Market Risk Contagion Amplifiers via Graph Attention Networks: Empirical Evidence from US Financial Stress Periods. *Journal of Computing Innovations and Applications, 4*(1), 164-175.

[34] Chen, Y., & Hu, J. (2026). Graph Neural Network-Based Cascading Disruption Path Identification in Multi-Tier Rare Earth Processing Networks. *Journal of Global Engineering Review, 4*(1), 99-112.

[35] Chen, Y. (2024). Explainable Attack Path Reasoning for Industrial Control Network Security Based on Knowledge Graphs. *Journal of Computing Innovations and Applications, 2*(1), 128-139.

[36] Hu, J., & Long, X. (2024). Graph Learning-Based Behavioral Detection for Software Supply Chain Attacks. *Journal of Advanced Computing Systems, 4*(4), 49-60.

[37] Wu, C., & Pan, Z. (2024). An Integrated Graph Neural Network and Reinforcement Learning Framework for Intelligent Drug Discovery. *Journal of Advanced Computing Systems, 4*(6), 19-29.

[38] Wei, C., & Wu, C. (2024). Credit Risk Transmission Mechanism and Prevention Strategies in Supply Chain Finance: A Core Enterprise Perspective. *Artificial Intelligence and Machine Learning Review, 5*(2), 101-115.

[39] Peng, Y., & Gao, S. (2026). An Empirical Comparison of XGBoost and LightGBM for Capital Expenditure Deviation Prediction in US FERC-Regulated Rate-Base Transmission Projects. *Journal of Sustainability, Policy, and Practice, 2*(4), 90-102.

[40] Hu, J., Wang, X., & Lai, J. (2026). Benchmarking Learned Cardinality Estimation Techniques for Analytical Query Processing in Data Warehouses. *Journal of Computer Technology and Applied Mathematics, 3*(3), 1-8.

[41] Chen, Y., & Chen, Z. (2025). Multi-Objective Deep Reinforcement Learning for Carbon-Aware Spatiotemporal Workload Scheduling in Geo-Distributed Data Centers. *Journal of Advanced Computing Systems, 5*(10), 18-30.

[42] Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling. *Artificial Intelligence and Machine Learning Review, 6*(4), 18-31.

[43] Long, L., & Hu, J. (2026). Multi-Objective Particle Swarm Optimization for Site Selection and Policy Subsidy Maximization of Foreign Renewable Energy Enterprises in the United States. *Artificial Intelligence and Machine Learning Review, 7*(2), 54-69.

[44] 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

[45] Zhu, S., Dekker, R., van Jaarsveld, W., Renjie, R. W., & Koning, A. J. (2017). An improved method for forecasting spare parts demand using extreme value theory. *European Journal of Operational Research, 261*(1), 169–181.

[46] Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 International Conference on Prognostics and Health Management* (pp. 1–9). IEEE. https://doi.org/10.1109/PHM.2008.4711414

[47] Arias Chao, M., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data, 6*(1), 5. https://doi.org/10.3390/data6010005

[48] Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1913643

[49] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146–3154).

[50] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31* (pp. 6638–6648).

[51] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. https://doi.org/10.1198/016214506000001437

[52] Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

[53] Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Tsetlin, I., & Winkler, R. L. (2022). The M5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting, 38*(4), 1365–1385. https://doi.org/10.1016/j.ijforecast.2021.10.009

[54] Wang, S., Kang, Y., & Petropoulos, F. (2024). Combining probabilistic forecasts of intermittent demand. *European Journal of Operational Research, 315*(3), 1038–1048.

[55] Kang, A., Li, Z., & Meng, S. (2023). AI-Enhanced Risk Identification and Intelligence Sharing Framework for Anti-Money Laundering in Cross-Border Income Swap Transactions. Journal of Advanced Computing Systems 3 (5), 34-47.

[56] Kang, A., Zhang, K., & Chen, Y. (2025). AI-Assisted Analysis of Policy Communication during Economic Crises: Correlations with Market Confidence and Recovery Outcomes. Pinnacle Academic Press Proceedings Series 3, 159-173.

[57] Kang, A., Xin, J., & Ma, X. (2024). Anomalous cross-border capital flow patterns and their implications for national economic security: An empirical analysis. Journal of Advanced Computing Systems 4 (5), 42-54.

[58] Kang, A., & Ma, X. (2025). AI-Based Pattern Recognition and Characteristic Analysis of Cross-Border Money Laundering Behaviors in Digital Currency Transactions. Pinnacle Academic Press Proceedings Series 5, 1-19.

[59] Kang, A., & Yu, K. (2025). The Impact of Financial Data Visualization Techniques on Enhancing Budget Transparency in Local Government Decision-Making. Spectrum of Research 5 (2).

[60] Wang, Z., & Kang, A. (2025). FTAFO: A Federated Transparent Adaptive Financial Optimizer for Reducing Third-Party Dependencies in Workflow Management. Journal of Science, Innovation & Social Impact 1 (1), 329-339.

[61] Kang, A., Li, C., & Meng, S. (2025). The Impact of Government Budget Data Visualization on Public Financial Literacy and Civic Engagement. Journal of Economic Theory and Business Management 2 (4), 1-16.

[62] Kang, A., Min, S., & Yuan, D. (2024). Comparative Analysis of Foreign Exchange Market Shock Transmission and Recovery Resilience Among Major Economies Under Geopolitical Conflicts: Evidence from the Russia-Ukraine…. Journal of Computing Innovations and Applications 2 (1), 46-61.

[63] Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems 4 (2), 36-49.

[64] Zhang, D., & Ma, X. (2025). Machine Learning-Based Credit Risk Assessment for Green Bonds: Climate Factor Integration and Default Prediction Analysis. Journal of Sustainability, Policy, and Practice 1 (2), 121-135.

[65] Yuan, D., & Zhang, D. (2025). APAC-sensitive anomaly detection: Culturally-aware AI models for enhanced AML in US securities trading. 2025 International Conference on Computer, AI, Systems and Automation, 108-121.

[66] Dong, B., Zhang, D., & Xin, J. (2024). Deep reinforcement learning for optimizing order book imbalance-based high-frequency trading strategies. Journal of Computing Innovations and Applications 2 (2), 33-43.

[67] Zhang, D., & Feng, E. (2024). Quantitative Assessment of Regional Carbon Neutrality Policy Synergies Based on Deep Learning. Journal of Advanced Computing Systems 4 (10), 38-54.

[68] Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems 5 (7), 1-17.

[69] Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science 3 (5), 27-40.

[70] Zhang, D., & Wang, Y. (2025). AI-driven quality assessment and investment risk identification for carbon credit projects in developing countries. 2025 International Forum on Computing and Intelligent Systems, 76-92.

[71] Rao, G. L., Trinh, T. K., Chen, Y., Shu, M., & Zheng, S. (2024). Jump Prediction in Systemically Important Financial Institutions' CDS Prices. Spectrum of Research 4 (2).

[72] Rao, G. L., Lu, T., Yan, L., & Liu, Y. (2024). A Hybrid LSTM-KNN Framework for Detecting Market Microstructure Anomalies:: Evidence from High-Frequency Jump Behaviors in Credit Default Swap Markets. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online ….

[73] Rao, G. L., Wang, Z., & Liang, J. (2025). Reinforcement Learning for Pattern Recognition in Cross-Border Financial Transaction Anomalies: A Behavioral Economics Approach to AML. Applied and Computational Engineering 142, 116-127.

[74] Rao, G., Ju, C., & Feng, Z. (2024). AI-Driven Identification of Critical Dependencies in US-China Technology Supply Chains: Implications for Economic Security Policy. Journal of Advanced Computing Systems 4 (12), 43-57.

[75] Ge, L., & Rao, G. (2025). MultiStream-FinBERT: A Hybrid Deep Learning Framework for Corporate Financial Distress Prediction Integrating Accounting Metrics, Market Signals, and Textual Disclosures. Pinnacle Academic Press Proceedings Series 3, 107-122.

[76] Rao, G., Zheng, S., & Guo, L. (2025). Dynamic Reinforcement Learning for Suspicious Fund Flow Detection: A Multi-layer Transaction Network Approach with Adaptive Strategy Optimization. Applied and Computational Engineering 145, 1-11.

[77] Ju, C., & Rao, G. (2025). Analyzing Foreign Investment Patterns in the US Semiconductor Value Chain Using AI-Enabled Analytics: A Framework for Economic Security. Pinnacle Academic Press Proceedings Series 2, 60-74.

[78] Liu, W., Rao, G., & Lian, H. (2023). Anomaly Pattern Recognition and Risk Control in High-Frequency Trading Using Reinforcement Learning. Journal of Computing Innovations and Applications 1 (2), 47-58.

[79] Wu, Z., Wang, S., Ni, C., & Wu, J. (2024). Adaptive traffic signal timing optimization using deep reinforcement learning in urban networks. Artificial Intelligence and Machine Learning Review 5 (4), 55-68.

[80] Xiong, K., Wu, Z., & Jia, X. (2025). Deepcontainer: a deep learning-based framework for real-time anomaly detection in cloud-native container environments. Journal of Advanced Computing Systems 5 (1), 1-17.

[81] Wu, Z., Feng, E., & Zhang, Z. (2024). Temporal-Contextual Behavioral Analytics for Proactive Cloud Security Threat Detection. Academia Nexus Journal 3 (2).

[82] Lei, Y., & Wu, Z. (2025). A Real-Time Detection Framework for High-Risk Content on Short Video Platforms Based on Heterogeneous Feature Fusion. 2025 International Forum on Computing and Intelligent Systems, 93-106.

[83] Wu, Z., Zhang, Z., Zhao, Q., & Yan, L. (2025). Privacy-preserving financial transaction pattern recognition: A differential privacy approach. Preprints.

[84] Zhang, Z., & Wu, Z. (2023). Context-aware feature selection for user behavior analytics in zero-trust environments. Journal of Advanced Computing Systems 3 (5), 21-33.

[85] Wu, Z., Feng, Z., & Dong, B. (2024). Optimal feature selection for market risk assessment: A dimensional reduction approach in quantitative finance. Journal of Computing Innovations and Applications 2 (1), 20-31.

[86] Wu, Z., Cheng, C., & Zhang, C. (2025). Cloud-Enabled AI Analytics for Urban Green Space Optimization: Enhancing Microclimate Benefits in High-Density Urban Areas. 2025 International Forum on Computing and Intelligent Systems, 123-133.

[87] Li, Y., Wu, C., Li, L., Liu, Y., & Zhu, J. (2021). Caption generation from road images for traffic scene modeling. IEEE Transactions on intelligent transportation systems 23 (7), 7805-7816.

[88] Li, L., Li, Y., Wu, C., Dong, H., Jiang, P., & Wang, F. (2021). Detail fusion GAN: High-quality translation for unpaired images with GAN-based data augmentation. 2020 25th International Conference on Pattern Recognition (ICPR), 1731-1736.

[89] Huo, H., Li, Y., Wu, C., Wu, X., Tian, Z., & Liu, Y. (2019). Semantic segmentation and scene reconstruction for traffic simulation using CNN. 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence ….

[90] Wu, X., Li, Y., Hao, Z., Wu, C., Wang, X., & Liu, Y. (2019). Image style transformation based on structure GAN. 2019 Chinese Automation Congress (CAC), 2002-2007.

[91] Wu, C., Li, Y., Li, L., Wang, L., & Liu, Y. (2020). Caption Generation from Road Images for Traffic Scene Construction. 2020 IEEE Intelligent Vehicles Symposium (IV), 1271-1276.

[92] Wu, X., Li, Y., Liu, Y., Pang, S., Wang, L., Wu, C., & Huo, H. (2019). Jointly detecting and retrieving vehicles from road image sequences based on CNN. 2019 IEEE Intelligent Vehicles Symposium (IV), 2530-2535.