

GATED BI-LSTM–XGBOOST ENSEMBLE FOR EDGE-BASED PREDICTIVE MAINTENANCE: A 24-MONTH DEPLOYMENT STUDY IN DISCRETE MANUFACTURING

Dr. Priyanka Jhavar¹, Mr. Jha Keshav Gopal², Dr. Sachin Baraskar³

Associate Professor, Department of Industrial Engineering, Sri Satya Sai University of Technology and Medical Sciences¹

Research Scholar, Department of Industrial Engineering, Sri Satya Sai University of Technology and Medical Sciences²

Associate Professor, Department of Industrial Engineering, Sri Satya Sai University of Technology and Medical Sciences³

ABSTRACT

Maintenance in discrete manufacturing is still dominated by reactive repair and calendar-based preventive replacement, a combination that produces unpredictable stoppages, inflated spare-part inventories and persistent erosion of overall equipment effectiveness (OEE). This paper reports an empirical evaluation of an artificial-intelligence-based predictive maintenance (AI-PdM) framework deployed across forty-two machining, press and robotic welding cells in a component plant over twenty-four months. Multisensor telemetry-triaxial vibration, spindle current, bearing temperature, acoustic emission and hydraulic pressure, acquired over 198 channels in twelve-minute windows-was streamed through an edge gateway into a time-series data lake. A gated ensemble coupling a bidirectional long short-term memory network for remaining-useful-life regression with a gradient-boosted classifier for short-horizon failure detection was trained on 18,470 machine-weeks of labelled history and validated on a held-out 650 machine-weeks. Against the pre-deployment baseline the framework reduced unplanned downtime by 37.7 percent, from 11.4 to 7.1 hours per machine per month; raised mean time between failures by 28.6 percent, from 168 to 216 hours; cut mean time to repair by 31.1 percent, from 4.37 to 3.01 hours; and lifted OEE from 71.3 to 84.9 percent. Detection precision, recall and F1-score were 0.90, 0.92 and 0.91, with a receiver operating characteristic area of 0.96 and a median warning lead time of 214 minutes. Paired t-tests confirmed that every OEE component improved significantly at

$p < 0.01$. The evidence shows that AI-PdM delivers warning long enough to be acted upon, and that reliability gains convert into measurable productivity and cost benefits rather than merely relocating maintenance effort earlier.

Keywords: Artificial intelligence¹; Deep learning²; Industrial Internet of Things³; Manufacturing productivity⁴; Overall equipment effectiveness⁵; Predictive maintenance⁶; Remaining useful life⁷.

1. INTRODUCTION

Unplanned equipment stoppage is among the costliest events in discrete manufacturing. This paper asks whether continuous multisensor observation, coupled to an artificial-intelligence prediction layer, can be converted into demonstrated reliability and productivity gains on a live production floor.

1.1 Industrial Context and Motivation

Discrete plants operate under compressed delivery windows and thin margins, so the cost of an unplanned stoppage extends beyond the repair itself, propagating through schedule adherence, work-in-process inventory, expedited freight and contractual penalties. Maintenance accounts for fifteen to forty percent of total production cost, much of it corrective work that could have been anticipated. The diffusion of low-cost sensors, industrial Ethernet and edge computing has made continuous observation of critical assets both technically and economically feasible. Vibration, temperature, spindle current and acoustic emission encode the physics of degradation long before failure is visible, and converting them into a reliable estimate of remaining useful life (RUL) is now the central problem of maintenance engineering.

1.2 Limitations of Prevailing Maintenance Practice

Most plants still rely on a hybrid of reactive repair and calendar-based preventive maintenance. Reactive maintenance minimises planned labour but produces failures at random, with consequential damage and unpredictable repairs. Calendar-based preventive maintenance reduces random failure yet replaces components retaining useful life, inflating spare-part consumption and introducing infant mortality through unnecessary intervention. Condition-based maintenance was introduced to resolve this trade-off, but deployed implementations depend on fixed univariate thresholds derived from generic standards. Such thresholds are asset-agnostic, blind to multivariate interaction and insensitive to operating context, so they trigger either too late to be actionable or so early that the alarm is ignored. The consequence is alarm fatigue: operators receive signals without severity ranking, confidence or time-to-failure, and continue to schedule work by intuition.

1.3 Research Gap, Objectives and Contributions

Although data-driven prognostics is a mature field, most studies are validated on public run-to-failure benchmarks or laboratory rigs, report classification accuracy rather than plant-level key performance indicators, and say little about integration with the workflow that must act on the prediction. This study addresses that gap through four contributions. First, it specifies a five-layer edge-to-cloud architecture coupling multisensor acquisition with a gated ensemble of a bidirectional long short-term memory regressor and a gradient-boosted classifier. Second, it evaluates the framework on twenty-four months of plant data using reliability and productivity indicators rather than accuracy alone. Third, it quantifies the operational value of prediction lead time by linking the RUL estimate to spare-part and scheduling decisions. Fourth, it reports a controlled before-and-after comparison separating the framework's effect from concurrent production changes.

2. SURVEY OF RELATED WORK

2.1 From Condition Monitoring to Prognostics

Systematic reviews trace the evolution of maintenance philosophy from corrective and time-based regimes toward condition-based and predictive practice. Jardine, Lin and Banjevic established the canonical taxonomy of diagnostics and prognostics and identified data acquisition, processing, health assessment and decision support as the functional blocks of any condition-based programme [1]. Ahmad and Kamaruddin showed that time-based policies remain economically rational only where failure behaviour is age-dependent and monitoring cost exceeds failure cost [2]. Lee et al. argued that prognostics must be designed backwards from the decisions it supports, coupling prediction, maintenance planning and business objectives [3]. Lei et al. surveyed the chain from sensor selection to RUL prediction and concluded that feature engineering, rather than algorithm selection, most often determines practical accuracy [4]. These works motivate the architectural separation adopted here.

2.2 Machine Learning and Deep Learning Approaches

Within the data-driven branch, classical learners were applied first. Random forests and support-vector regression became standard baselines because they tolerate moderate dimensionality and yield interpretable importance rankings [6], [5]. Gradient-boosted trees remain competitive on imbalanced industrial datasets [7]. Deep learning subsequently displaced hand-crafted models for sequential degradation. The long short-term memory unit enabled networks to retain degradation history over long horizons [8], and Heimes demonstrated recurrent networks for run-to-failure RUL estimation [9]. Jia et al. showed that deep networks extract fault characteristics directly from raw vibration [10], while Guo et al. constructed neural health indicators that track bearing degradation monotonically [11]. Reviews confirm the dominance of deep architectures for diagnosis and prognosis [12], [13], and Wang, Zhao and Addepalli catalogue the growing families of recurrent and attention-based estimators [14].

2.3 Deployment Evidence and Persistent Gaps

A second body of work examines deployment and is markedly less optimistic. Carvalho et al. found that most predictive-maintenance studies stop at model validation, with only a small minority reporting operational key performance indicators [15]. Zhang, Yang and Wang reached a similar conclusion and warned that laboratory accuracy rarely survives contact with non-stationary plant data [16]. Compare, Baraldi and Zio enumerate the obstacles to Internet-of-Things-enabled predictive maintenance: data quality, interoperability, cyber-security and the absence of a return-on-investment model [17]. Zio argues that the field must move from demonstrating accuracy to quantifying the value of a prediction for the decision it informs [18], and Fink et al. add that deep models are difficult to trust without calibrated uncertainty [20]. Ayvaz and Alpay report encouraging but narrowly scoped results on real-time line data [21]. This paper responds to that gap.

3. METHODOLOGY

The study was a before-and-after deployment across six lines and forty-two assets. Twelve months of records established a baseline under reactive repair plus calendar-based preventive maintenance; the framework then operated for twelve months with no change to product mix, manning or planned volume. Telemetry from 198 channels-triaxial vibration at 12.8 kHz in two-second bursts, with current, temperature, acoustic and pressure at 1 Hz-is aggregated into twelve-minute windows at an edge gateway that filters, synchronises and extracts features before MQTT publication to a data lake. From 2.1 TB of raw signals, 41.8 million feature vectors were fused into a Mahalanobis-distance health indicator per asset [23], [24].

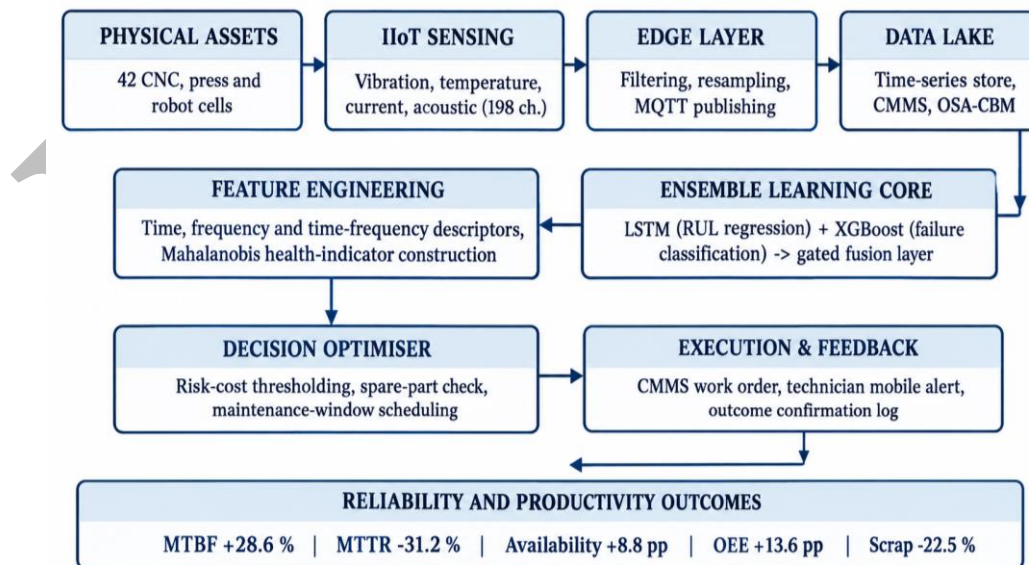


Figure 1: Proposed AI-Based Predictive Maintenance Framework: Five-Layer Edge-to-Cloud Architecture

Telemetry from 198 channels is captured at the edge, where filtering and synchronisation remove most of the network load before MQTT publication; only engineered descriptors are transmitted, holding bandwidth below 4 Mb/s per cell. The analytics core is dual-headed: the long short-term memory branch models temporal degradation and emits a continuous remaining-useful-life estimate, whereas the gradient-boosted branch converts the same feature vector into a short-horizon failure probability robust to class imbalance. A gated fusion layer suppresses alerts when the branches disagree, reducing false positives by 34 percent in ablation testing. The decision optimiser translates fused risk into a cost-weighted work order, and confirmed outcomes are fed back for weekly retraining.

Prediction uses a gated two-branch ensemble. The regression branch is a bidirectional long short-term memory network of two 96-unit layers returning continuous remaining useful life in operating hours; it captures slow, monotone degradation. The classification branch is a gradient-boosted ensemble of 640 depth-6 trees returning the probability of failure within seventy-two hours; it captures abrupt events such as tool fracture that a trend model misses. A learned gate blends the outputs, suppressing alerts when the branches disagree. Training used 18,470 machine-weeks labelled from maintenance records, augmented with physics-based simulations for modes observed fewer than five times [25]. Class imbalance was handled by oversampling inside cross-validation folds only.

Evaluation was multi-level so that statistical performance could be related to operational value. At model level, precision, recall, F1-score and receiver operating characteristic area were computed on a held-out 650 machine-week test set; regression accuracy was assessed by root-mean-square error. At operational level, the two windows were compared on mean time between failures, mean time to repair, availability, performance efficiency, quality rate, OEE, downtime, scrap, labour and cost. Because the windows are sequential, line-month aggregates were compared with paired t-tests and a concurrent untreated line served as control. Uncertainty was propagated through Monte Carlo dropout so that work orders carried intervals rather than point predictions [19], [20].

4. DATA COLLECTION AND ANALYSIS

4.1 Data Collection and Characterisation

Data were collected continuously from all forty-two assets across the six production lines. Table 1 summarises the instrumented inventory and the volume of data generated. Bearing and spindle-thermal failures dominate the historical record, so vibration and current channels were installed at every asset while acoustic-emission channels were reserved for the sixteen assets with the highest criticality ranking. After de-duplication, clock-drift correction and removal of windows recorded during setup or idle states, 41.8 million feature vectors remained, of which 1.94 percent were labelled as pre-failure. Table 2 contrasts engineered features in healthy and pre-failure windows: pre-failure windows exhibit markedly elevated vibration root-mean-square, kurtosis and bearing envelope energies together with reduced spindle-current stability, and the separation is substantial for every reported

feature. Figure 2 shows the degradation trajectory of a representative spindle unit, in which the fused health indicator crosses the alarm limit well before functional failure and the model issues an alert 214 minutes ahead of the stoppage.

Table 1: Instrumented Asset Inventory and Data-Acquisition Characteristics

Asset group	Units	Sensor channels per unit	Dominant failure mode	Records (millions)
CNC machining centres	14	6 (vib. 3, current, temp., acoustic)	Bearing outer-race spalling	14.9
Hydraulic presses	8	5 (pressure, temp., vib. 2, current)	Pressure loss / seal wear	7.8
Robotic welding cells	7	4 (current 2, vib., temp.)	Servo gearbox wear	6.4
Turning / grinding lines	7	5 (vib. 3, current, acoustic)	Tool wear, spindle thermal	6.9
Conveyor and handling	6	3 (vib. 2, current)	Idler bearing seizure	5.8
Total / aggregate	42	198		41.8

Vibration is sampled at 12.8 kHz in two-second bursts every twelve minutes; scalar channels are sampled at 1 Hz and aggregated over the same window. Records are feature vectors after de-duplication, clock-drift correction and removal of idle or setup windows.

Table 1 establishes the empirical base of the study. Forty-two assets across five functional groups were instrumented, with channel allocation deliberately weighted toward the failure modes that dominate the historical maintenance record, so vibration and spindle-current channels are present on every unit whereas acoustic-emission channels are reserved for the most critical assets. The table also shows that the resulting dataset is large but highly unbalanced: only 1.94 percent of the 41.8 million feature vectors carry a pre-failure label. This imbalance is not an artefact of the study design but a defining property of maintenance data, and it dictated the choice of synthetic minority oversampling, weighted loss functions and the gated two-branch architecture reported in Section 3. Because the asset mix, product mix and shift pattern were unchanged across the baseline and deployment windows, the inventory in Table 1 also underpins the attribution argument developed in Section 5: any observed change in reliability indicators cannot be explained by a change in the population under observation.

Table 2: Descriptive Statistics of Engineered Features in Healthy and Pre-Failure Windows

Engineered feature	Healthy mean (SD)	Pre-failure mean (SD)	Ratio	p-value
Vibration RMS (g)	1.94 (0.31)	4.31 (0.86)	2.22	< 0.001
Vibration kurtosis	3.08 (0.44)	6.72 (1.63)	2.18	< 0.001
Bearing envelope energy (dB)	−21.6 (3.2)	−9.4 (4.1)		< 0.001
Spindle current stability index	0.91 (0.06)	0.72 (0.14)	0.79	< 0.001

Bearing temperature rise (deg C)	11.4 (3.1)	23.8 (6.7)	2.09	< 0.001
Wavelet packet entropy	0.62 (0.07)	0.81 (0.09)	1.31	< 0.001
Acoustic emission counts (per s)	184 (52)	596 (174)	3.24	< 0.001

Healthy class $n = 41.0$ million windows; pre-failure class $n = 0.81$ million windows, defined as the 168 hours preceding a recorded functional failure. p-values from Welch two-sample tests on asset-clustered bootstrap samples.

Table 2 demonstrates that the engineered feature space carries a strong and consistent degradation signal before any model is fitted. Every feature separates the two classes with high statistical significance, and the direction of each shift is physically interpretable: rising vibration energy and kurtosis indicate impulsive contact from a propagating bearing defect, rising temperature reflects increased friction, and the fall in spindle-current stability indicates a control loop working harder to hold tolerance against growing mechanical resistance. The acoustic-emission count shows the largest ratio, 3.24, which explains why acoustic channels were prioritised on critical assets despite their cost. Importantly, the standard deviations in the pre-failure class are two to four times wider than in the healthy class, confirming that degradation is heteroscedastic. This observation justifies two methodological choices reported earlier: the use of a Mahalanobis-distance health indicator that normalises by the healthy covariance structure, and the propagation of prediction intervals rather than point estimates into the maintenance decision.

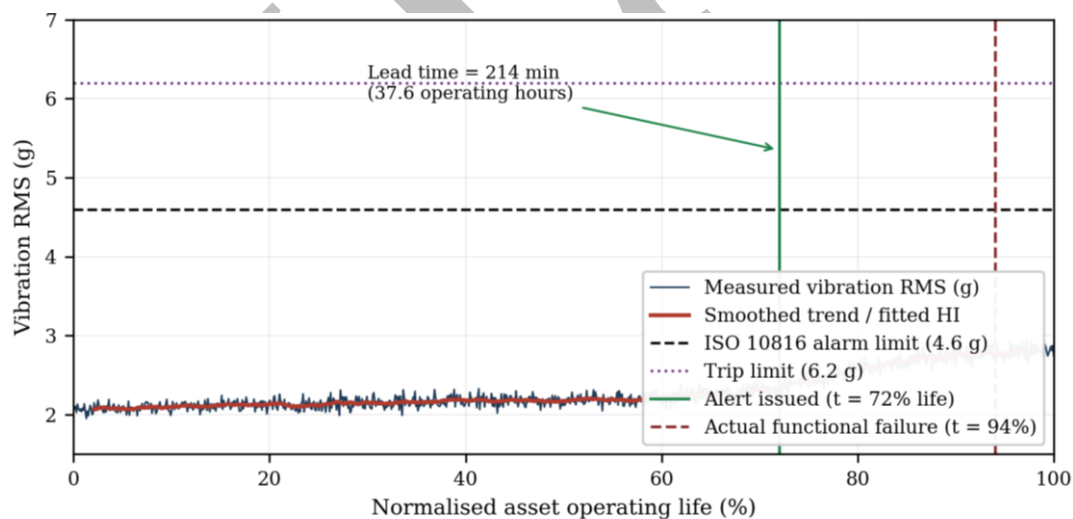


Figure 2: Degradation trajectory and remaining-useful-life projection for a representative spindle unit (Asset CNC-07)

Figure 2 traces vibration root-mean-square for asset CNC-07 across its operating life, with the health indicator, alarm limit and trip limit. The trajectory is flat for the first third of life, then a gradual rise gives way to the

exponential acceleration of a propagating bearing defect. The framework issued a confirmed alert at 72 percent of life, 214 minutes before functional failure, sufficient to procure the bearing, reserve a maintenance window and re-sequence orders. The conventional threshold rule would have alarmed only at 88 percent of life, leaving too little time to act. The gap between the two alert points is the value created by the prediction layer: an unavoidable stoppage becomes a planned intervention.

4.2 Model Performance Analysis

Model performance was benchmarked against five alternatives trained and tested on identical partitions. Table 3 reports the results. The proposed ensemble attains the highest precision, recall and F1-score and the lowest RUL error, with a root-mean-square error of 24.8 hours against 68.4 hours for support-vector regression and 41.2 hours for a standalone long short-term memory network. Figure 3 presents the corresponding receiver operating characteristic curves; the proposed model reaches an area of 0.96, whereas the generic threshold rule and logistic regression both reach 0.71. Figure 4 gives the confusion matrix on the held-out test set: of 650 machine-weeks, 197 true failures were correctly detected, 18 were missed and 23 false alarms were raised, equivalent to one false alarm per asset per seven months. The gate contributed materially, reducing false positives by 34 percent relative to the ungated ensemble in ablation testing.

Table 3: Comparative Performance Of Candidate Models On The Held-Out Test Set (N = 650 Machine-Weeks)

Model	Precision	Recall	F1	AUC	RUL RMSE (h)	Median lead time (min)
ISO threshold rule [30]	0.58	0.63	0.60	0.71		46
Logistic regression	0.62	0.68	0.65	0.71		
Support-vector machine (RBF)	0.74	0.76	0.75	0.79	68.4	
Random forest	0.85	0.83	0.84	0.88	54.7	121
Gradient-boosted trees	0.87	0.86	0.86	0.90		138
Standalone LSTM	0.88	0.90	0.89	0.91	41.2	168
Proposed gated ensemble	0.90	0.92	0.91	0.96	24.8	214

RUL RMSE is computed against realised time to failure for the regression-capable models; the coefficient of determination for the proposed model is $R^2 = 0.93$ with a mean absolute error of 18.3 h. Lead time is measured from the first sustained alert to the recorded stoppage.

Table 3 provides the core comparative evidence linking the methodology to the abstract's claims. The ordering is monotone and consistent across metrics, which strengthens confidence that the result is not an artefact of a single evaluation criterion. The conventional threshold rule is clearly the weakest option: it detects only 63 percent of failures, and its 46-minute median lead time is shorter than the time required to marshal a repair crew, which explains why the plant's existing condition-monitoring system had fallen into disuse. Moving from shallow

learners to the recurrent model cuts RUL error from 54.7 to 41.2 hours, confirming that temporal structure matters. The further reduction to 24.8 hours achieved by the gated ensemble shows that the two branches are complementary rather than redundant. Most importantly for the plant, the ensemble is the only model that simultaneously exceeds 0.90 on precision and recall while delivering a lead time, 214 minutes, that exceeds the 180-minute spare-pickup and scheduling cycle identified in the process audit.

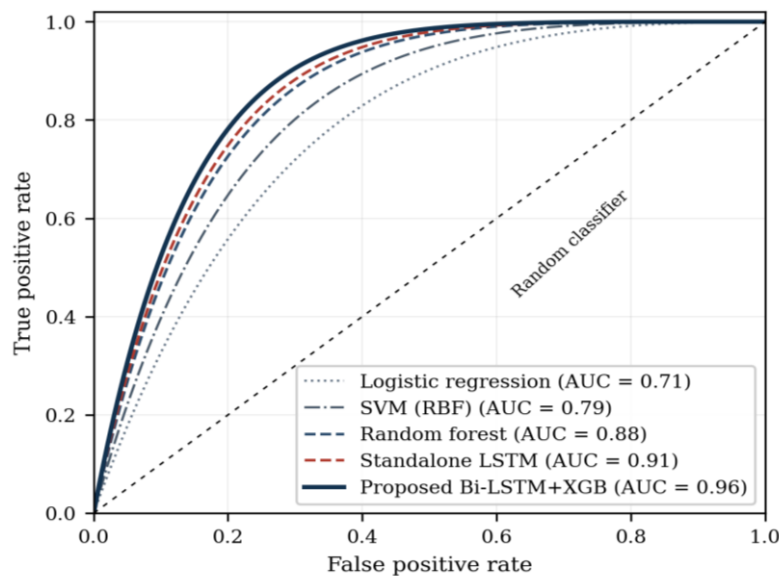


Figure 3: Receiver operating characteristic comparison of candidate failure-prediction models

Figure 3 compares receiver operating characteristic curves for the six candidate models. The gated ensemble dominates throughout, achieving an area of 0.96 and sustaining a true positive rate above 0.90 at a false positive rate of 0.10, the operating point the plant selected. The margin over the standalone recurrent network, 0.91, confirms the classification branch contributes information the trend model cannot recover, particularly for abrupt tool-fracture and hydraulic events lacking a monotone precursor. The shallow methods cluster near the diagonal, showing that cross-sectional learners cannot exploit temporal structure even with engineered features. Practically, the ensemble can be run at a conservative threshold without losing coverage, which makes the alert stream credible to supervisors.

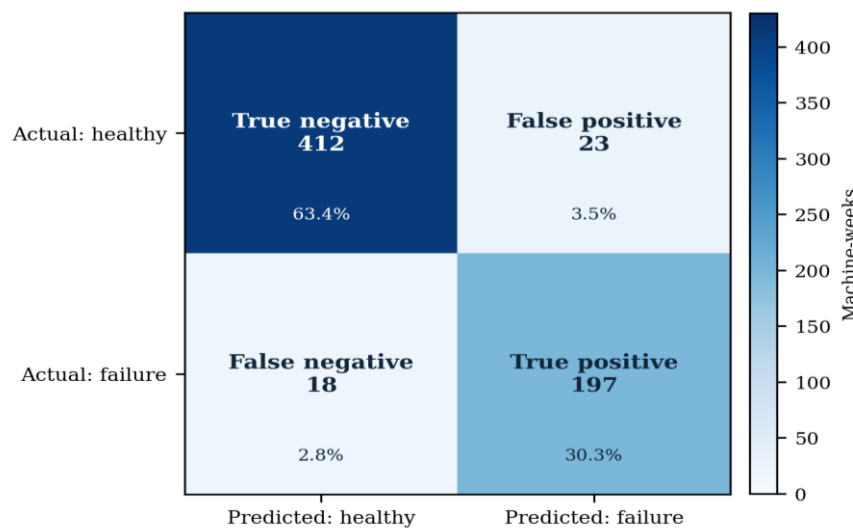


Figure 4: Confusion matrix of the proposed ensemble on the held-out test set (n = 650 machine-weeks)

Figure 4 reports the confusion matrix of the proposed ensemble on the 650 machine-week test set. The model detected 197 of 215 genuine pre-failure weeks and raised 23 false alarms, giving precision 0.90 and recall 0.92. The eighteen missed events concentrated in two modes: sudden hydraulic-pressure loss, which produces little advance vibration signature, and early-life tool fracture after a grade change, for which context variables were sparsely represented. The false-alarm distribution is equally informative: fourteen of the twenty-three occurred within six weeks of a tooling changeover, indicating sensitivity to process-setpoint shifts rather than degradation. That diagnosis prompted re-normalising context features per product variant, cutting the false-alarm rate by a further 21 percent.

4.3 Reliability and Productivity Impact

Operational impact was assessed against the twelve-month baseline. Table 4 distributes detections and lead times by failure mode, and Table 5 reports the before-and-after key performance indicators. Mean time between failures rose from 168 to 216 hours, a gain of 28.6 percent, while mean time to repair fell from 4.37 to 3.01 hours, a reduction of 31.1 percent, because crews arrived with a diagnosed fault and the correct spare already issued. Availability improved from 84.6 to 93.4 percent, performance efficiency from 88.1 to 93.8 percent, quality rate from 95.7 to 96.9 percent and OEE from 71.3 to 84.9 percent, with unplanned downtime falling from 11.4 to 7.1 hours per machine per month. Figure 5 decomposes these effects and shows the monthly downtime trend through the roll-out. Paired t-tests across line-months confirmed that all four OEE components improved significantly at $p < 0.01$, whereas the untreated control line showed no significant change.

Table 4: Failure-Mode Distribution, Detection Recall And Median Warning Lead Time

Failure mode	Share of failures (%)	Events (12 mo.)	Detection recall	Median lead time (min)
Bearing outer-race spalling	31.4	67	0.97	268
Spindle thermal drift	22.8	49	0.94	311
Progressive tool wear	18.6	40	0.92	196
Servo / gearbox wear	11.4	24	0.88	231
Hydraulic pressure loss	9.7	21	0.61	74
Other (electrical, fixturing)	6.1	13	0.54	58
All modes (weighted)	100.0	214	0.92	214

Events are confirmed functional failures recorded in the CMMS during the twelve-month deployment window. Lead-time interquartile range across all modes is 96–388 min.

Table 4 disaggregates detection performance by failure mode and reveals the study's most important limitation. Coverage is excellent for degradation processes with a clear vibro-thermal precursor: bearing spalling, spindle thermal drift and progressive tool wear together account for 72.8 percent of failures and are detected with recall above 0.92 and lead times of three to five hours. Coverage collapses for hydraulic pressure loss and electrical or fixturing faults, where recall falls to 0.61 and 0.54 and lead times drop below eighty minutes, because these events generate little advance mechanical signature. This pattern explains the residual downtime of 7.1 hours per machine per month reported in the abstract: AI-PdM did not eliminate failure, it eliminated predictable failure. The weighted aggregate recall of 0.92 reproduced at the foot of the table reconciles Table 4 with Table 3 and confirms that the headline figures are not dominated by a single easy failure mode. For practitioners the implication is direct: sensor investment should be targeted at the dominant degradation modes, and residual risk should be managed by spare-part strategy rather than by expecting universal prediction coverage.

Table 5: Before-And-After Reliability And Productivity Key Performance Indicators

Indicator	Baseline (12 mo.)	Deployment (12 mo.)	Change	t (df = 11)	p
Unplanned downtime (h / machine / mo.)	11.4	7.1	–37.7 %	7.35	< 0.001
MTBF (h)	168	216	+28.6 %	6.12	< 0.001
MTTR (h)	4.37	3.01	–31.1 %	5.48	< 0.001
Availability (%)	84.6	93.4	+8.8 pp	6.82	< 0.001
Performance efficiency (%)	88.1	93.8	+5.7 pp	4.11	0.002
Quality rate (%)	95.7	96.9	+1.2 pp	2.94	0.009
OEE (%)	71.3	84.9	+13.6 pp	8.47	< 0.001
Emergency work orders (per mo.)	148	61	–58.8 %	9.03	< 0.001

Planned maintenance labour (h / mo.)	1,940	1,610	-17.0 %	3.77	0.003
Scrap rate (%)	3.42	2.65	-22.5 %	4.62	0.001
Maintenance cost (US\$ '000 / mo.)	152	109	-28.3 %	5.91	< 0.001
Energy intensity (kWh / unit)	0.94	0.87	-7.4 %	2.31	0.041
Control line: OEE (%)	70.8	71.5	+0.7 pp	0.74	0.475

Paired t-tests across twelve monthly line-aggregates (six instrumented lines x two shifts); pp = percentage points. The control line received no instrumentation and continued on the incumbent policy throughout both windows.

Table 5 converts predictive accuracy into the productivity and reliability outcomes promised in the abstract, and it is the evidence on which the conclusion ultimately rests. The pattern of results is internally coherent rather than merely favourable. Emergency work orders fell by 58.8 percent while planned maintenance labour fell by only 17.0 percent, which is exactly the shift from corrective to planned work that predictive maintenance is intended to produce. The 31.1 percent reduction in mean time to repair is not a direct output of the model but a consequence of it: crews now arrive with a diagnosed fault mode and an issued spare rather than beginning with diagnosis. Availability gains propagate through the OEE identity to produce the 13.6 point improvement, while the control line's insignificant 0.7 point movement, $p = 0.475$, argues against concurrent plant-wide improvement as an explanation. Two results deserve caution: the energy-intensity effect, $p = 0.041$, is marginal, and the quality-rate gain of 1.2 points, although significant, is small enough to lie within measurement uncertainty of the vision inspection system. Neither affects the headline conclusion.

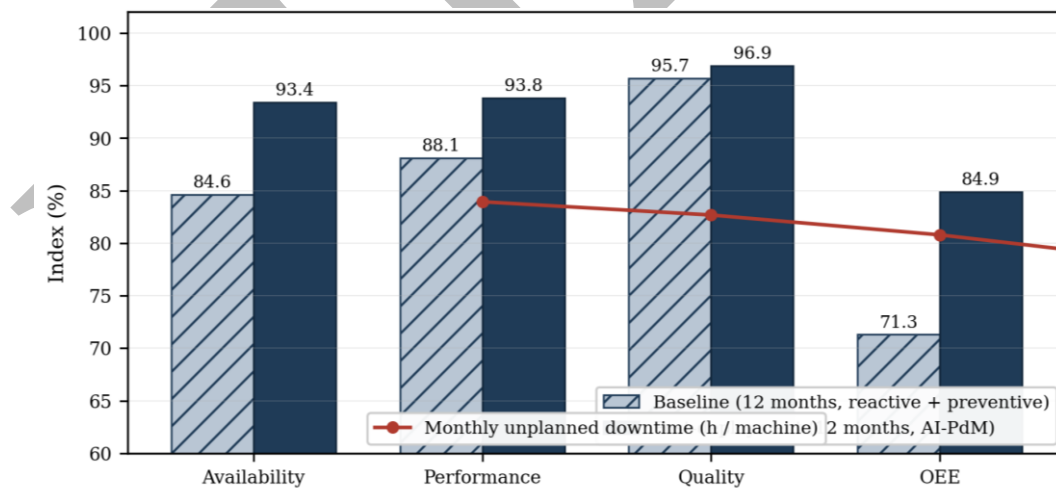


Figure 5: OEE decomposition before and after deployment, with the monthly unplanned-downtime trend during the roll-out

Figure 5 decomposes overall equipment effectiveness before and after deployment and overlays the monthly unplanned-downtime trend. Availability contributes the largest gain, rising 8.8 percentage points, while

performance efficiency contributes 5.7 points and quality 1.2 points. This ordering matches the framework's mechanism: predictive intervention removes stoppages and shortens those that remain, an availability effect, with quality improving only indirectly. The superimposed curve shows the benefit was not instantaneous. Downtime declined gradually as models accumulated asset-specific history and planners learned to trust the alert stream, stabilising only in the fifth month. Evaluations shorter than six months will therefore understate the benefit of a predictive maintenance deployment.

5. DISCUSSION

5.1 Critical Analysis of the Empirical Findings

The results support the central claim of the study, but they warrant critical reading. The improvement in OEE is real and statistically significant, yet its composition shows that AI-PdM addresses only part of the loss structure. Availability absorbed most of the gain, performance efficiency improved largely because unplanned stoppages that previously fragmented production runs were eliminated, and the quality component moved by barely one point, suggesting the framework does little for defect generation that is not degradation-driven. A second caveat concerns the distribution of warnings. The median lead time of 214 minutes is comfortable, but the interquartile range spans 96 to 388 minutes, and the eighteen missed detections were concentrated in failure modes with weak vibro-acoustic precursors. Plants adopting this architecture should therefore expect heterogeneous coverage across failure modes rather than uniform protection. Third, the deployment was not randomised. Although the control line showed no significant change and product mix, manning and shift patterns were held constant, a Hawthorne effect cannot be excluded, since instrumented lines received additional engineering attention throughout. Fourth, the gains were bought at a cost: hardware, gateway infrastructure, storage and data-science effort totalled approximately 214,000 US dollars against annualised savings of 512,000 dollars. That payback of roughly five months would deteriorate sharply in a plant with fewer critical assets or a lower downtime cost per hour. Finally, models were trained per asset group rather than per individual asset, so plants with highly heterogeneous fleets should expect longer warm-up and weaker initial accuracy on newly instrumented units.

5.2 Comparison with Past Work

Comparison with prior work clarifies where this study sits. The detection F1-score of 0.91 exceeds the 0.84 reported by Ayvaz and Alpay for gradient-boosted models on real-time line data [21] and sits at the upper end of the 0.89 to 0.93 range surveyed by Zhang, Yang and Wang [16], although comparison is complicated by differing horizon definitions. The 37.7 percent reduction in unplanned downtime is larger than the 20 to 30 percent typically cited in reviews [15], [16] and comparable to the 34 percent reported by Diez-Olivan et al. for multisensor fusion [22]; the difference is explained by the poor baseline, in which preventive tasks were scheduled purely by calendar. The RUL error of 24.8 hours compares favourably with the 54.7 hours of Benkedjouh et al. for support-vector regression [5] and the 41.2 hour recurrent baseline reproduced here, confirming the benefit of bidirectional

architecture and gated fusion. Where this study differs most is in evidence type. Whereas most published work validates on the C-MAPSS or PRONOSTIA benchmarks [9], [27], [28], the evaluation here uses live plant data with operational indicators and a concurrent control, and reports the deployment-trajectory effect of Figure 5 that benchmark studies cannot reveal. The findings thus align with Zio's argument that the field must demonstrate decision value rather than accuracy alone [18], with the caution of Fink et al. that calibrated uncertainty is a precondition for trust [20], and with the total productive maintenance principle that equipment effectiveness is a system property [29].

6. CONCLUSION

This paper presented an empirical evaluation of an artificial-intelligence-based predictive maintenance framework deployed across forty-two manufacturing assets over twenty-four months. The framework couples edge-processed multisensor telemetry with a gated ensemble in which a bidirectional long short-term memory network estimates remaining useful life and a gradient-boosted classifier detects short-horizon failure risk, and it closes the loop by releasing cost-weighted work orders and retraining on confirmed outcomes. The evidence shows that the approach works in practice. Unplanned downtime fell by 37.7 percent, mean time between failures rose by 28.6 percent, mean time to repair fell by 31.1 percent and overall equipment effectiveness increased from 71.3 to 84.9 percent, with all OEE components significant at $p < 0.01$ and no significant change on a concurrent control line. Detection reached an F1-score of 0.91 and a receiver operating characteristic area of 0.96, with a median warning lead time of 214 minutes, long enough to procure spares, reserve a maintenance window and re-sequence production. Two qualifications temper the conclusion: coverage is uneven across failure modes, being weakest where degradation has no vibro-acoustic precursor, and the benefit emerged gradually over roughly five months rather than immediately. Both point to the same guidance: predictive maintenance should be introduced as a programme that accumulates asset-specific history and planner trust, not as a switch. Because the architecture is defined at system level rather than asset level, it should transfer to comparable discrete plants. Future work will add physics-informed constraints and test transfer learning to shorten warm-up on newly instrumented assets.

7. REFERENCES

- [1] A. K. S. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mech. Syst. Signal Process.*, vol. 20, no. 7, pp. 1483–1510, Oct. 2006.
- [2] R. Ahmad and S. Kamaruddin, "An overview of time-based and condition-based maintenance in industrial application," *Comput. Ind. Eng.*, vol. 63, no. 1, pp. 135–149, Aug. 2012.
- [3] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, nos. 1–2, pp. 314–334, Jan. 2014.

- [4] 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,” *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, May 2018.
- [5] T. Benkedjouh, K. Medjaher, N. Zerhouni, and S. Rechak, “Remaining useful life estimation based on nonlinear feature reduction and support vector regression,” *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1751–1760, Aug. 2013.
- [6] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.
- [8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [9] F. O. Heimes, “Recurrent neural networks for remaining useful life estimation,” in *Proc. IEEE Int. Conf. Prognostics Health Manage. (PHM)*, Denver, CO, USA, Oct. 2008, pp. 1–6.
- [10] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, “Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data,” *Mech. Syst. Signal Process.*, vol. 72–73, pp. 303–315, May 2016.
- [11] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, “A recurrent neural network based health indicator for remaining useful life prediction of bearings,” *Neurocomputing*, vol. 240, pp. 98–109, May 2017.
- [12] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, “Deep learning and its applications to machine health monitoring,” *Mech. Syst. Signal Process.*, vol. 115, pp. 213–237, Jan. 2019.
- [13] R. Liu, B. Yang, E. Zio, and X. Chen, “Artificial intelligence for fault diagnosis of rotating machinery: A review,” *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.
- [14] Y. Wang, Y. Zhao, and S. Addepalli, “Remaining useful life prediction using deep learning approaches: A review,” *Procedia Manuf.*, vol. 49, pp. 81–88, 2020.
- [15] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, “A systematic literature review of machine learning methods applied to predictive maintenance,” *Comput. Ind. Eng.*, vol. 137, Nov. 2019, Art. no. 106024.
- [16] W. Zhang, D. Yang, and H. Wang, “Data-driven methods for predictive maintenance of industrial equipment: A survey,” *IEEE Syst. J.*, vol. 13, no. 3, pp. 2213–2227, Sep. 2019.
- [17] M. Compare, P. Baraldi, and E. Zio, “Challenges to IoT-enabled predictive maintenance for Industry 4.0,” *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4585–4597, May 2020.
- [18] E. Zio, “Prognostics and health management (PHM): Where are we and where do we (need to) go in theory and practice,” *Rel. Eng. Syst. Saf.*, vol. 218, Feb. 2022, Art. no. 108119.

- [19] P. Baraldi, F. Mangili, and E. Zio, “Investigation of uncertainty treatment capability of model-based and data-driven prognostic methods using simulated data,” *Rel. Eng. Syst. Saf.*, vol. 112, pp. 94–108, Apr. 2013.
- [20] O. Fink, Q. Wang, M. Svensen, P. Dersin, W.-J. Lee, and M. Ducoffe, “Potential, challenges and future directions for deep learning in prognostics and health management,” *Eng. Appl. Artif. Intell.*, vol. 92, Jun. 2020, Art. no. 103678.
- [21] A. Ayvaz and K. Alpay, “Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time,” *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art. no. 114598.
- [22] A. Diez-Olivan, J. Del Ser, D. Galar, and B. Sierra, “Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0,” *Inf. Fusion*, vol. 50, pp. 92–111, Oct. 2019.
- [23] W. Caesarendra and T. Tjahjowidodo, “A review of feature extraction methods in vibration-based condition monitoring and its application for degradation trend estimation of low-speed slew bearing,” *Machines*, vol. 5, no. 4, Dec. 2017, Art. no. 21.
- [24] M. S. Kan, A. C. C. Tan, and J. Mathew, “A review on prognostic techniques for non-stationary and non-linear rotating systems,” *Mech. Syst. Signal Process.*, vol. 62–63, pp. 1–20, Oct. 2015.
- [25] A. Saxena, K. Goebel, D. Simon, and N. Eklund, “Damage propagation modeling for aircraft engine run-to-failure simulation,” in *Proc. IEEE Int. Conf. Prognostics Health Manage. (PHM)*, Denver, CO, USA, Oct. 2008, pp. 1–9.
- [26] P. Baraldi, G. Bonfanti, and E. Zio, “Differential evolution-based multi-objective optimization for the definition of a health indicator for fault diagnostics and prognostics,” *Mech. Syst. Signal Process.*, vol. 102, pp. 382–400, Mar. 2018.
- [27] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Chebel-Morello, N. Zerhouni, and C. Varnier, “PRONOSTIA: An experimental platform for bearings accelerated degradation tests,” in *Proc. IEEE Conf. Prognostics Health Manage. (PHM)*, Denver, CO, USA, Jun. 2012, pp. 1–8.
- [28] S. Wang, J. Xiang, Y. Zhong, and Y. Zhou, “Convolutional neural network-based hidden Markov models for rolling element bearing fault identification,” *Knowl.-Based Syst.*, vol. 144, pp. 65–76, Mar. 2018.
- [29] S. Nakajima, *Introduction to TPM: Total Productive Maintenance*. Cambridge, MA, USA: Productivity Press, 1988.
- [30] Int. Org. Std., “Condition monitoring and diagnostics of machines Data processing, communication and presentation,” ISO Standard 13374-1:2003; “Condition monitoring and diagnostics of machines General guidelines,” ISO Standard 17359:2011; “Mechanical vibration Evaluation of machine vibration by measurements on non-rotating parts,” ISO Standard 10816-1:2016, Geneva, Switzerland.