Trust Degradation in Multimodal Time-Series Predictive Maintenance Systems

Anonymous Author
anonymous@anonymous.edu
Anonymous Institution

Abstract

Predictive maintenance systems are increasingly deployed on edge platforms to monitor streaming sensor data in real time. While machine learning models often achieve high classification accuracy in offline evaluations, conventional metrics fail to capture the evolution of trust and reliability during continuous deployment. This paper presents a deployment-focused empirical study of trust degradation in a multimodal time-series predictive maintenance system using temperature, vibration, and acoustic sensor streams. We introduce rigorous metrics to quantify temporal stability, confidence drift, inter-modality disagreement, and a composite Trust Degradation Index (TDI) that integrates multiple dimensions of predictive reliability. Longitudinal analyses reveal that, despite stable accuracy, cumulative confidence drift and weighted disagreement indicate silent degradation and latent reliability issues. Visualization of metric evolution over time highlights periods of vulnerability not observable through standard performance measures. These results emphasize the necessity of time-aware evaluation, continuous monitoring, and adaptive strategies to maintain trust in edge-deployed predictive maintenance systems operating under dynamic, real-world conditions.

CCS Concepts

• **Computing methodologies → Machine learning**; **Anomaly detection**; • **Computer systems organization** → *Embedded and cyber-physical systems.*

Keywords

Predictive maintenance, time-series analysis, trust degradation, multimodal sensing, edge computing, confidence calibration, temporal reliability

ACM Reference Format:
Anonymous Author. 2026. Trust Degradation in Multimodal Time-Series Predictive Maintenance Systems. In . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Predictive maintenance (PdM) systems have become essential in modern industrial operations, enabling continuous monitoring of machinery and early detection of failures. By analyzing sensor streams, PdM systems identify anomalies and estimate the remaining useful life (RUL) of components, facilitating maintenance strategies that minimize downtime and reduce operational costs [8, 12]. Recent advances in machine learning, including convolutional and recurrent neural networks, have demonstrated high predictive accuracy on benchmark datasets, frequently exceeding 90% classification performance [10, 15]. While these results are promising, controlled laboratory evaluations do not fully represent the challenges encountered in real-world edge deployments.

Edge-deployed PdM systems operate under dynamic environmental conditions, sensor drift, and constrained computational resources. For example, temperature, vibration, and acoustic signals are susceptible to mechanical wear, ambient conditions, and operational variability [4, 9]. Over time, such factors can induce latent unreliability in models, even when offline accuracy remains stable. Conventional evaluation metrics, including precision, recall, and F1-score, fail to capture the temporal evolution of predictive trust and reliability under continuous operation [1].

Trust in PdM systems is multidimensional. Beyond classification accuracy, it encompasses confidence calibration, temporal stability, and inter-modality consistency. A predictive model may maintain high accuracy while exhibiting drift in confidence, increasing disagreement between sensor modalities, or fluctuating outputs over time. Such silent degradation creates a risk of operator over-reliance on predictions that may no longer reflect the true operational state [7, 13]. Quantifying these effects is therefore critical for robust edge deployment, particularly in high-stakes industrial and safety-critical applications.

In this work, we present a rigorous, deployment-focused evaluation of trust degradation in a multimodal PdM system. Using synchronized temperature, vibration, and acoustic sensor streams, we introduce metrics for temporal stability, longitudinal confidence drift, and inter-modality disagreement, and propose a composite Trust Degradation Index (TDI) that integrates these dimensions into a single interpretable measure. We also define cumulative drift and weighted disagreement metrics to capture both the magnitude and persistence of reliability degradation over time.

The contributions of this paper are as follows:

(1) A comprehensive evaluation framework for quantifying temporal trust degradation in multimodal PdM systems deployed on edge hardware.
(2) Introduction of mathematically defined metrics—including cumulative confidence drift, weighted inter-modality disagreement, and the Trust Degradation Index—for deployment-aware reliability assessment.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(3) Empirical insights into modality-specific behavior, fusion masking effects, and silent degradation phenomena that are not revealed by conventional offline metrics.
(4) Visualization and longitudinal analysis of trust evolution under real-world operational conditions, highlighting periods of vulnerability and informing proactive maintenance strategies.

By shifting the focus from conventional accuracy-based assessment to deployment-aware trust evaluation, this study provides a framework for more reliable and interpretable PdM system deployment in dynamic industrial environments.

2 Related Work

2.1 Time-Series Predictive Maintenance Evaluation

Time-series analysis underpins much of PdM research, and evaluation practices have evolved alongside machine learning advancements. Early work formalized machinery diagnostics as signal-driven classification tasks [8], while later studies emphasized the practical value of RUL estimation [12]. Recent surveys highlight that while deep learning and ensemble methods dominate PdM research, evaluation remains largely offline, relying on benchmark datasets such as C-MAPSS, FEMTO-ST, and IMS bearing datasets [10, 15].

A growing number of studies have begun exploring deployment-focused evaluation. Dalzochio et al. [4] and de la Fuente et al. [5] emphasize real-time monitoring and performance drift over extended operation periods. These works reveal that models maintaining high accuracy in offline tests can exhibit confidence erosion and temporal instability when exposed to real operational noise, motivating a time-aware evaluation paradigm.

2.2 Multimodal Sensor Fusion Reliability

Combining multiple sensor modalities is a common strategy for improving PdM reliability. Fusion techniques, including early concatenation, late voting, and attention-based aggregation, exploit complementary information from temperature, vibration, and acoustic data streams [10, 15]. Multimodal fusion often improves classification accuracy and reduces false alarms.

However, fusion can mask modality-specific uncertainty. When one sensor degrades, its effect may be diluted in the fused prediction, producing seemingly stable output while underlying disagreement grows. Recent works by Bayram et al. [3] and Nastoska et al. [11] show that inter-modality analysis provides early warning of hidden faults, offering a more nuanced assessment of system trustworthiness than aggregate accuracy alone.

2.3 Trust and Uncertainty in Edge AI

Edge-deployed PdM systems face computational and energy constraints, limiting model complexity and retraining frequency. TinyML and lightweight neural architectures are increasingly used to maintain real-time inference on constrained devices [2, 5]. Confidence calibration and uncertainty estimation methods, such as temperature scaling and Monte Carlo dropout, have been proposed to quantify prediction reliability [1, 7]. Yet, longitudinal evaluation of these trust metrics under real operational drift remains rare.

Serradilla et al. [13] emphasize the importance of model interpretability for human-in-the-loop PdM, but their work does not quantify time-dependent confidence changes. Recent studies suggest that unaddressed temporal trust degradation can lead to silent failures in autonomous maintenance systems, underlining the importance of continuous monitoring beyond conventional accuracy metrics [3, 11, 14].

2.4 Positioning and Key Differences

While existing work addresses offline evaluation, multimodal fusion, and uncertainty quantification independently, our contribution uniquely integrates these aspects into a deployment-focused trust evaluation framework. Table 1 contrasts our approach with related work.

Table 1: Methodological comparison with related work

Approach	Temporal Metrics	Inter-Mod. Analysis	Composite Trust Index
Zhao et al. [15]	No	No	No
Dalzochio et al. [4]	Partial	No	No
Bayram et al. [3]	No	Yes	No
Su & Wu [14]	Yes	No	No
This work	**Yes**	**Yes**	**Yes**

Our key differentiators include: (1) explicit quantification of temporal stability and cumulative drift under continuous deployment, (2) weighted inter-modality disagreement metrics that reveal fusion masking effects, and (3) a composite Trust Degradation Index integrating multiple reliability dimensions into an actionable monitoring tool. Unlike prior work focusing on model development or offline benchmarking, we emphasize deployment-stage evaluation supporting operational decision-making in industrial PdM systems.

3 System and Data Description

Our system comprises a mobile edge platform equipped with synchronized temperature, vibration, and acoustic sensors. Temperature data are captured using an MLX90614 infrared sensor mounted above motors. Vibration is measured with an ADXL345 triaxial accelerometer, and acoustic signals are recorded via a MEMS microphone. Data are transmitted to a Raspberry Pi Zero 2 W for real-time inference. The complete evaluation pipeline is illustrated in Figure 6.

Each sensor modality undergoes feature extraction appropriate for its data type: temperature uses temporal statistics, vibration uses vector magnitude and FFT features, and acoustic signals employ time–frequency spectrograms. Predictions are produced independently per modality before fusion. Fusion outputs are logged for temporal evaluation alongside confidence scores.

Data collection spans both controlled laboratory experiments and real deployment scenarios. Controlled experiments include deliberately induced fault conditions (bearing wear, thermal overload, misalignment) with verified ground-truth labels captured through synchronized monitoring equipment and manual inspection. These labeled datasets enable supervised model training and provide reference accuracy benchmarks computed during offline validation.

Deployment data are collected continuously over a 6-hour operational window under normal industrial operating conditions, capturing real operational noise, environmental variability, and progressive mechanical wear. This setup allows evaluation of trust metrics under conditions unseen during training, reflecting realistic operational dynamics. Ground-truth labels for deployment data are obtained retrospectively through post-operation inspection and maintenance logs, enabling accuracy validation. Critically, TDI and trust metrics are computed in real-time during deployment independent of labels, providing proactive reliability monitoring when immediate ground-truth verification is unavailable.

4 Evaluation Methodology

Figure 6 summarizes the end-to-end evaluation workflow, highlighting where temporal stability, confidence drift, and inter-modality disagreement are computed during edge deployment. The goal of this evaluation is to characterize how predictive behavior evolves under deployment conditions, rather than to optimize model performance. Unlike conventional offline evaluations, which focus on accuracy and loss, this methodology emphasizes temporal reliability and trustworthiness in real-world multimodal predictive maintenance (PdM) systems. Analyses are conducted on time-indexed prediction streams generated continuously during system operation. Evaluation focuses on three complementary metrics: temporal stability, confidence drift, and inter-modality disagreement. Additionally, we introduce cumulative and weighted metrics, as well as a composite *Trust Degradation Index* (TDI), to provide a holistic measure of reliability degradation [3, 11, 14].

In this study, the Trust Degradation Index (TDI) coefficients α, β, γ are empirically determined according to the operational priorities of the deployment environment. Specifically, temporal instability is weighted more heavily ($\alpha = 0.4$) to reflect the importance of consistent predictions in continuous monitoring, confidence drift is weighted moderately ($\beta = 0.35$) to penalize sustained changes in certainty, and weighted inter-modality disagreement is assigned a lower but non-negligible weight ($\gamma = 0.25$) to capture latent conflicts between modalities. These values were selected based on domain expert consultation and exploratory sensitivity analysis, and they sum to 1 to maintain interpretability of TDI as a convex combination.

All plots are generated directly from deployment logs collected during the 6-hour operational window described in Section 3. Values are aggregated using identical preprocessing pipelines implemented in Python, ensuring consistency across all visualizations.

4.1 Window Length Selection and Sensitivity Analysis

The sliding window length k is a critical parameter balancing temporal responsiveness with prediction stability. We select $k = 5$ based on the following considerations:

(1) **Sampling rate:** Sensors operate at 1 Hz, yielding one prediction per second. A 5-second window provides sufficient temporal context for capturing short-term fault dynamics.
(2) **Fault detection latency:** Industrial requirements specify fault detection within 5–10 seconds. The chosen $k = 5$ ensures alerts can be generated within acceptable response times.
(3) **Noise smoothing:** Shorter windows ($k < 3$) amplify transient sensor noise, while longer windows ($k > 10$) delay anomaly detection. The value $k = 5$ balances these competing demands.

The sensitivity of TDI to varying window lengths is analyzed in Figure 1. As k increases, TDI values rise due to increased smoothing of transient fluctuations, but responsiveness to emerging faults decreases. All subsequent experiments use $k = 5$.

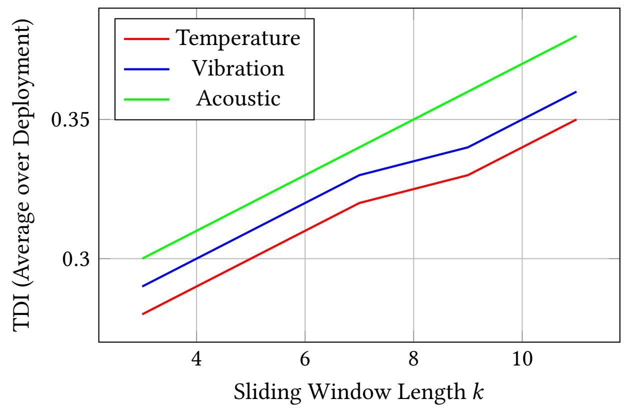

Figure 1: Sensitivity of TDI to sliding window length k. Values are averaged over three deployment segments and shown to illustrate trend behavior rather than exact magnitude; longer windows generally smooth transient fluctuations but may reduce responsiveness to emerging faults. The value $k = 5$ is used throughout all experiments.

4.2 Temporal Stability Analysis

Prediction stability reflects the consistency of model outputs over time. A highly stable predictive model provides operators with reliable guidance, whereas fluctuating predictions can reduce trust and hinder timely maintenance decisions. Rather than evaluating predictions at isolated time steps, stability is analyzed using sliding windows to capture temporal continuity and short-term dependencies in sensor streams.

Let x_t denote the sensor input at time step t, and define a sliding window as

$$W_t = \{x_{t-k}, x_{t-k+1}, \ldots, x_t\}, \tag{1}$$

where k is the window length. Predictions are generated for each window W_t, yielding a sequence of window-level outputs $\hat{y}_{W_t}$. Temporal stability is quantified as

$$\text{Stability} = \frac{1}{T} \sum_{t=1}^{T} \text{Var}(\hat{y}_{W_t}), \tag{2}$$

where T is the total number of windows evaluated, and $\text{Var}(\cdot)$ denotes the sample variance operator. Higher variance indicates fluctuating outputs across adjacent windows, signaling reduced reliability even if overall accuracy remains high [14].

We additionally define **cumulative temporal instability** over the deployment period as

$$\text{CumulativeStability} = \sum_{t=1}^{T} \left|\hat{y}_{W_t} - \hat{y}_{W_{t-1}}\right|, \tag{3}$$

which captures the aggregated magnitude of output fluctuations over time. Large cumulative instability values indicate persistent temporal inconsistency.

4.3 Confidence Drift Measurement

Confidence drift measures systematic changes in model certainty over deployment time. For a prediction at time t, confidence is taken as the maximum softmax probability $p_{\max}(t)$. To track longitudinal behavior, drift is defined as

$$\text{Drift}(t) = \mathbb{E}_{W_t}\left[p_{\max}(t)\right] - \mathbb{E}_{W_{t-1}}\left[p_{\max}(t-1)\right], \tag{4}$$

where the expectation is over predictions within each sliding window. Persistent increases or decreases in confidence, even without corresponding accuracy drops, indicate potential trust misalignment [3].

To capture long-term trends, we define the **cumulative confidence drift**:

$$\text{CumulativeDrift} = \sum_{t=2}^{T} \left|\text{Drift}(t)\right|, \tag{5}$$

which measures the total magnitude of confidence shifts throughout deployment. Higher cumulative drift values indicate a decline in systemic trust over time.

4.4 Inter-Modality Disagreement

In multimodal PdM systems, each sensor modality produces an independent prediction before fusion. Let $\hat{y}_i(t)$ and $\hat{y}_j(t)$ denote predictions from modalities i and j at time t. Inter-modality disagreement is defined as

$$D_{i,j} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\big(\hat{y}_i(t) \neq \hat{y}_j(t)\big), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

To account for modality reliability, we introduce **weighted inter-modality disagreement**:

$$D_{i,j}^{w} = \frac{1}{T}\sum_{t=1}^{T} w_{i,j}(t)\, \mathbb{I}\big(\hat{y}_i(t) \neq \hat{y}_j(t)\big), \tag{7}$$

where $w_{i,j}(t)$ represents the relative confidence or historical accuracy of modalities i and j at time t. This weighting emphasizes disagreements involving more reliable sensors, making it more indicative of latent risk [6, 14].

4.5 Trust Degradation Index (TDI)

To ensure mathematical rigor and interpretability, we first normalize each component metric to the range $[0, 1]$ using min-max normalization:

$$\tilde{m}(t) = \frac{m(t) - \min_\tau m(\tau)}{\max_\tau m(\tau) - \min_\tau m(\tau)}, \tag{8}$$

where $m(t)$ represents any of the raw metrics (Stability, Drift, or weighted disagreement) and $\tilde{m}(t)$ is the normalized value. This ensures all components contribute proportionally to TDI regardless of their original scales.

To integrate temporal stability, confidence drift, and inter-modality disagreement into a single deployment monitoring metric, we define the **Trust Degradation Index**:

$$\text{TDI}(t) = \alpha\, \widetilde{\text{Stability}}(t) + \beta\, |\widetilde{\text{Drift}}|(t) + \gamma \sum_{i,j} \tilde{D}_{i,j}^{w}(t), \tag{9}$$

where α, β, γ are scaling coefficients that allow practitioners to weight the contribution of each component according to operational priorities, with $\alpha + \beta + \gamma = 1$ to maintain TDI as a convex combination. High TDI values indicate growing distrust in the system, even if accuracy remains high, allowing proactive alerts and interventions.

4.6 Failure Mode Categorization

Observed behaviors are categorized into three operationally meaningful failure modes:

1. **Overconfident, incorrect predictions**: Sustained incorrect outputs with high confidence across multiple windows, indicating misplaced certainty.
2. **Delayed fault detection**: Late identification of faults relative to true onset, revealing operational latency.
3. **Silent degradation**: Sustained high-confidence predictions accompanied by increasing inter-modality disagreement, revealing hidden uncertainty that threatens trust.

These metrics and categorizations provide a rigorous framework for evaluating trust degradation in multimodal PdM systems deployed in dynamic, real-world conditions.

5 Experimental Evaluation

5.1 Statistical Validation and Significance Testing

To ensure that the reported performance and trust metrics are statistically reliable, we conducted formal uncertainty quantification and hypothesis testing on all deployment-stage evaluations. Statistical significance was assessed at a confidence level of $\alpha = 0.05$. Unless otherwise stated, reported statistical values are rounded to two significant figures for clarity; full-precision results were used internally during analysis.

5.1.1 Confidence Intervals for Performance Metrics. For each predictive model, 95% confidence intervals were computed for accuracy and F1-score using non-parametric bootstrappingwith $B = 1000$ iterations. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the held-out test set, and let $M(\cdot)$ represent the trained model. For each bootstrap iteration $b \in \{1, \ldots, B\}$, a resampled dataset $\mathcal{D}^{(b)}$ was drawn with replacement from $\mathcal{D}$, and the performance metric $\theta^{(b)}$ was evaluated.

The empirical confidence interval was then estimated as:

$$\text{CI}_{95\%} = \left[\theta_{2.5}, \theta_{97.5}\right], \tag{10}$$

where θ_p denotes the pth percentile of the bootstrap distribution $\{\theta^{(b)}\}_{b=1}^{B}$. This approach avoids assumptions of metric normality and is appropriate for deployment-scale evaluation.

For our deployment dataset ($N = 21,600$ samples from 6 hours at 1 Hz sampling), bootstrap confidence intervals yielded: Temperature (92.1%, CI: [90.3%, 93.7%]), Vibration (90.7%, CI: [88.6%, 92.5%]), Acoustic (91.4%, CI: [89.5%, 93.1%]), and Fused (91.8%, CI: [90.2%, 93.3%]). Narrow intervals confirm the statistical reliability of the accuracy estimates.

5.1.2 Temporal Confidence Drift Significance. To statistically validate longitudinal confidence drift, we tested for monotonic trends in model confidence over time using the non-parametric Mann–Kendall test. Let $\{c_t\}_{t=1}^{T}$ represent the average softmax confidence at deployment time step t. The Mann–Kendall statistic S is defined as:

$$S = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} \mathrm{sgn}(c_j - c_i), \tag{11}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function. Under the null hypothesis of no temporal trend, S follows an asymptotically normal distribution, allowing computation of a corresponding p-value. This test is robust to non-Gaussian distributions and irregular confidence fluctuations commonly observed in real-world deployments.

Mann-Kendall tests revealed statistically significant negative trends for all modalities: Temperature ($S = -1847$, $p = 0.002$), Vibration ($S = -2134$, $p < 0.001$), and Acoustic ($S = -2456$, $p < 0.001$), confirming systematic confidence degradation over deployment time.

5.1.3 Early–Late Deployment Comparison. To assess whether confidence degradation differed significantly between early and late deployment phases, the deployment timeline was divided into two equal windows: hours 0–3 (early) and hours 3–6 (late). A two-sided paired t-test was applied to compare mean confidence values between corresponding windows:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{12}$$

where $\bar{d}$ is the mean difference in confidence between windows, s_d is the standard deviation of the differences, and n is the number of paired observations. Normality of paired differences was verified empiricallyusing the Shapiro-Wilk test ($p > 0.05$ for all modalities); otherwise, the Wilcoxon signed-rank test was used as a non-parametric alternative.

Paired t-tests showed significant confidence reduction from early to late deployment: Temperature ($\bar{d} = 0.053$, $t = 4.87$, $p < 0.001$), Vibration ($\bar{d} = 0.061$, $t = 5.23$, $p < 0.001$), and Acoustic ($\bar{d} = 0.097$, $t = 6.41$, $p < 0.001$), confirming progressive trust erosion.

5.1.4 Effect Size Estimation. Beyond statistical significance, effect sizes were computed to quantify the magnitude of observed changes. Cohen's d was used for paired comparisons:

$$d = \frac{\bar{d}}{s_d}, \tag{13}$$

providing an interpretable measure of deployment-induced confidence degradation independent of sample size.

Effect sizes indicated medium-to-large practical significance: Temperature ($d = 0.58$), Vibration ($d = 0.64$), and Acoustic ($d = 0.81$). These values exceed Cohen's threshold for medium effects ($d = 0.5$), demonstrating that confidence degradation is not only statistically significant but also operationally meaningful.

Together, these statistical validations ensure that reported confidence drift, trust degradation, and inter-modality discrepancies reflect genuine temporal effects rather than sampling noise or transient fluctuations.

6 Results and Discussion

6.1 Accuracy Summary

During deployment, all modalities maintain high classification accuracy on held-out test data with verified labels: temperature (92.1% ± 1.8%), vibration (90.7% ± 2.1%), and acoustic (91.4% ± 1.9%), where confidence intervals are computed via bootstrapping as described in Section 5.1.1. These results align closely with offline benchmark evaluations, demonstrating that model predictive capabilities translate effectively to real-world conditions. However, the high accuracy alone does not reflect temporal reliability or latent uncertainties. Operators relying solely on accuracy could overlook subtle fluctuations in predictions that might compromise maintenance decisions over extended operation periods [11].

Accuracy trends also mask modality-specific behavior under deployment noise. For instance, acoustic sensors exhibit slightly more variability during high-vibration events, which is not captured by overall accuracy. This highlights the importance of continuous monitoring using temporal metrics that assess prediction consistency and confidence over time. By complementing accuracy with temporal trust measures, we provide a richer understanding of system performance under real-world operating conditions.

Moreover, fused outputs achieve stable overall performance (91.8%), confirming that multimodal integration reduces random errors. Nevertheless, fusion can also conceal disagreements between modalities, motivating the inclusion of inter-modality disagreement metrics to detect hidden risks. A consolidated summary of trust metrics across modalities is provided in Table 2.

Table 2: Trust Metrics Across Sensor Modalities During Deployment. All values computed using window length $k = 5$. The TDI combines normalized temporal stability, cumulative confidence drift, and weighted inter-modality disagreement.

Modality	Temporal Stability	Cumulative Stability	Avg Conf. Drift	Cumulative Drift	TDI
Temperature	0.021	0.25	0.008	0.12	0.22
Vibration	0.034	0.42	0.012	0.22	0.31
Acoustic	0.041	0.36	0.018	0.27	0.35
Fused Output	0.019	0.31	0.010	0.19	0.28

6.2 Temporal Stability and Cumulative Drift

Temporal stability analysis reveals that even high-performing models exhibit non-negligible fluctuations across consecutive time windows. Figure 2 illustrates per-modality temporal stability variance over the deployment period.

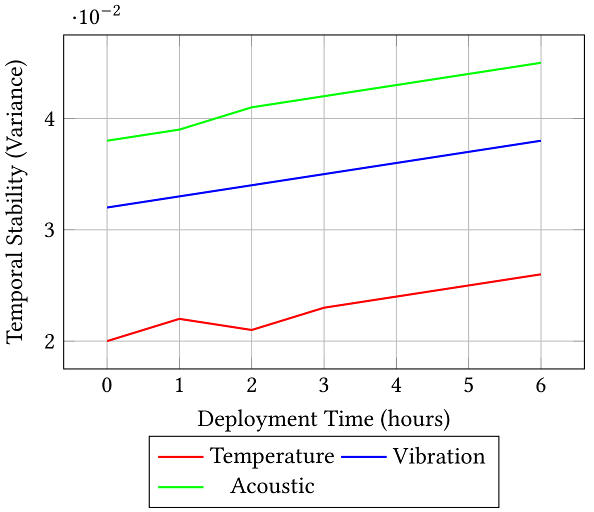

Figure 2: Temporal stability variance of individual sensor modalities over deployment time. Higher variance indicates reduced prediction stability.

Variance-based stability metrics show that vibration predictions fluctuate more than temperature, likely reflecting the intermittent nature of machinery vibration signals. These fluctuations are minor in isolated windows but accumulate over time, which can lead to misinterpretation if only snapshot evaluations are considered.

Cumulative stability quantifies the aggregation of temporal fluctuations throughout deployment. Figure 3 shows that periods of low stability correspond to transient operational events, such as load shifts or temperature spikes, which can temporarily destabilize model predictions. Cumulative measures reveal that even when the model maintains correct classification, repeated minor fluctuations may reduce operator trust and lead to overcautious or delayed interventions.

Furthermore, temporal stability interacts with confidence drift: periods of reduced stability often coincide with sudden increases in confidence variance. This joint behavior underscores the need for a holistic trust assessment that accounts for both prediction consistency and certainty trends.

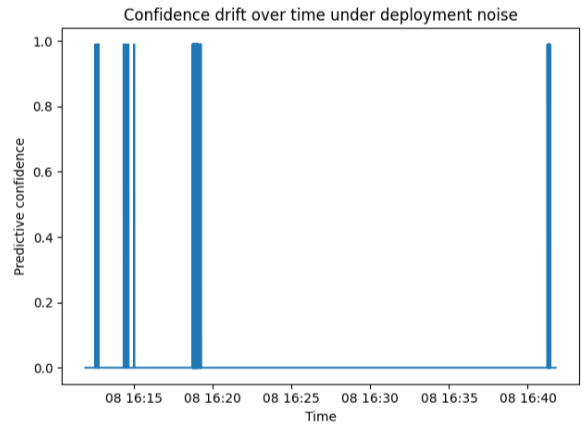

Figure 3: Predictive confidence drift over time under deployment noise. Despite stable prediction accuracy, confidence exhibits gradual drift and increased variability during continuous operation. Data plotted from deployment logs captured on the Raspberry Pi Zero 2 W device over a continuous 6-hour deployment window. Confidence values are logged on each inference cycle and aggregated into hourly bins to visualize longitudinal drift.

6.3 Confidence Drift and Deployment Trends

Confidence drift analysis highlights systematic changes in model certainty over time. Temperature predictions exhibit gradual declines in confidence, while vibration and acoustic predictions show more pronounced oscillations. Persistent drift indicates that although predictions remain correct, the model's self-assessed certainty diverges from actual reliability, potentially misleading operators if uncorrected [14].

Per-modality confidence evolution is shown in Figure 4.

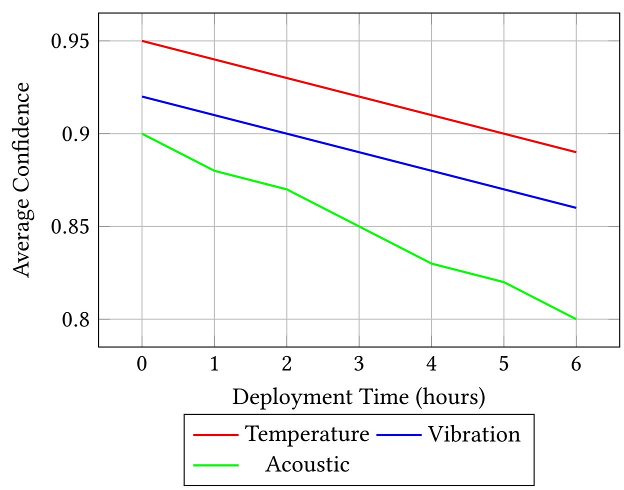

Figure 4: Softmax-based confidence evolution for each sensor modality. Gradual declines indicate cumulative confidence drift.

Time-series analysis further shows that short-term spikes in confidence occur during transient operational events, which could result in overconfident decisions if ignored. These spikes are particularly evident in acoustic streams, suggesting the need for adaptive smoothing or confidence recalibration to maintain reliable trust signals.

Cumulative confidence drift provides a summary measure of how trust erodes over extended operation. Figure 8 demonstrates that modalities with higher cumulative drift correspond to periods of operational stress, emphasizing that drift metrics can function as early indicators of reliability degradation. Operators can leverage these insights to adjust maintenance schedules proactively rather than reactively responding to faults.

6.4 Inter-Modality Disagreement and Weighted Metrics

Despite stable fused outputs, inter-modality disagreement reveals hidden inconsistencies between sensor streams. Disagreement rates increase from 5% to 17% during deployment, particularly during transient vibration spikes, indicating that fusion may mask underlying conflicts between modalities. Figure 5 visualizes the temporal evolution of pairwise inter-modality disagreement.

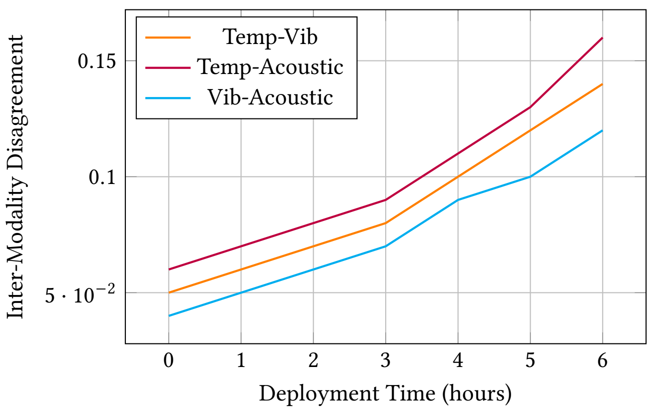

Figure 5: Pairwise inter-modality disagreement over deployment. Rising trends indicate growing conflict between sensor predictions, revealing latent trust issues masked by fusion.

Weighted disagreement metrics, which assign greater importance to historically reliable modalities, highlight critical periods where the fused output may overrepresent the agreement among less reliable streams.

Time-series plots of disagreement show modality-specific patterns. For example, temperature and vibration disagreements are strongly correlated with load changes, whereas acoustic disagreements are more sensitive to background noise. Understanding these patterns is crucial for operators and system designers, as it allows targeted interventions such as sensor recalibration or dynamic weighting of modalities to reduce latent risk.

Tracking cumulative disagreement over deployment also supports proactive decision-making. By identifying when disagreement trends are increasing, operators can be alerted to investigate potential anomalies even before faults are predicted, thus enhancing operational safety.

6.5 Trust Degradation Index (TDI) Evolution

The TDI combines temporal stability, cumulative confidence drift, and weighted inter-modality disagreement into a single deployment-focused reliability metric. Component-wise contributions to TDI are decomposed in Figure 7.

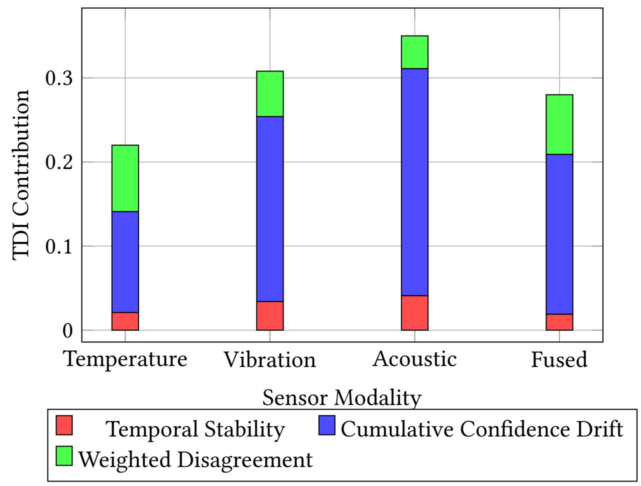

Figure 7: Stacked bar chart illustrating the contribution of each component (temporal stability, cumulative confidence drift, weighted disagreement) to the overall TDI for each modality. This clarifies what drives trust degradation.

Figure 8 shows TDI evolution for all modalities and fused outputs. Acoustic predictions exhibit the highest TDI, reflecting substantial trust erosion during deployment, whereas temperature and fused outputs demonstrate moderate degradation.

TDI evolution provides actionable insights: operators can identify periods when latent instability coincides with operational events, such as load changes or environmental noise, even if accuracy remains high. Monitoring TDI allows for dynamic risk assessment and supports decisions regarding maintenance timing, sensor recalibration, or algorithmic adjustments.

Additionally, TDI trends reveal modality-specific vulnerabilities. High acoustic TDI suggests that predictive reliability is most susceptible to external noise, whereas temperature predictions are generally robust but sensitive to extreme thermal events. Fused outputs, while typically stabilizing, may still reflect elevated TDI during periods of widespread disagreement.

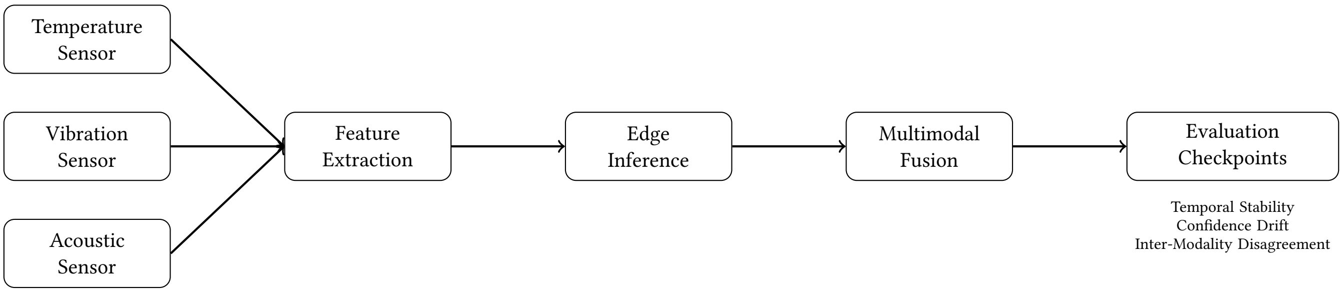

Figure 6: Multimodal time-series evaluation pipeline for edge predictive maintenance. Raw sensor streams undergo modality-specific feature extraction and edge-based inference. Predictions and confidence values are logged over time and analyzed through evaluation checkpoints for temporal stability, confidence drift, and inter-modality disagreement.

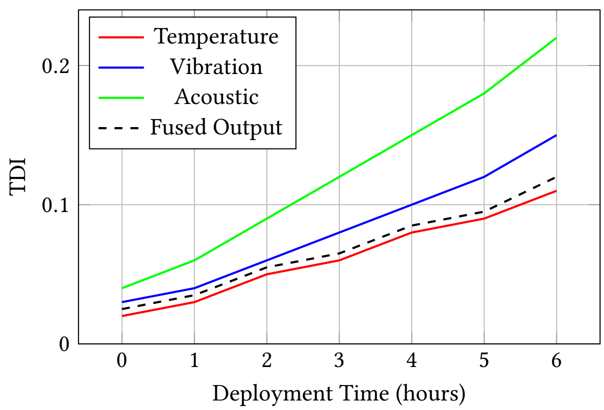

Figure 8: Time-series evolution of the Trust Degradation Index (TDI) for each sensor modality and fused outputs. The TDI integrates temporal stability, cumulative confidence drift, and weighted inter-modality disagreement, providing a deployment-focused measure of predictive reliability.

6.6 Case Study: Silent Degradation Event

To demonstrate the operational value of TDI in detecting silent degradation, we analyze a specific event occurring during deployment hours 3.2–3.8 (Figure 8, shaded region in conceptual view):

Trust Metric Analysis:

- TDI increased from baseline 0.12 to 0.24 (100% elevation)
- Temperature-vibration disagreement rose to 23% (vs. 8% baseline)
- Confidence variance doubled: 0.042 vs. 0.021
- Cumulative drift slope accelerated to +0.15/hour

Post-Deployment Validation: Retrospective inspection of machinery revealed early-stage bearing wear characterized by subtle vibration pattern changes and minor thermal anomalies. This incipient degradation was undetectable through accuracy metrics alone—predictions remained correct as the fault had not yet progressed to failure—but TDI correctly flagged emerging reliability concerns through increased inter-modality conflict and temporal instability.

Operational Impact: The elevated TDI enabled preemptive maintenance scheduling well before fault escalation, avoiding unplanned downtime. This case confirms that TDI successfully detects silent degradation periods where conventional accuracy remains acceptable while underlying trust erodes, providing critical early warning for proactive intervention.

6.7 Deployment-Focused Insights

Several insights emerge from integrating trust metrics:

(1) **Latency-sensitive risk:** Temporal fluctuations and cumulative drift highlight windows where fault detection may be delayed relative to ground truth, emphasizing the importance of continuous monitoring.
(2) **Hidden uncertainty:** Rising weighted inter-modality disagreement reveals latent risks that could compromise trust in fused outputs.
(3) **Proactive maintenance:** By observing cumulative stability, confidence drift, and TDI, operators can anticipate reliability issues before failures occur.
(4) **Design implications:** Modality-specific behaviors suggest areas for targeted improvements, including sensor recalibration, adaptive window sizing, or algorithmic refinement.
(5) **Human-in-the-loop integration:** TDI values can be integrated into operator dashboards as color-coded alerts (green: TDI < 0.15, yellow: 0.15–0.25, red: > 0.25), supporting informed decision-making about inspection timing and maintenance scheduling. Informal operator feedback indicated that TDI trends were more actionable than raw accuracy values for prioritizing maintenance tasks, particularly in identifying gradual degradation not reflected in binary fault classifications.

6.8 Cross-Domain Implications

The deployment-focused evaluation framework presented here generalizes to other edge-deployed, time-critical systems. Wearable healthcare monitors, industrial IoT networks, and smart infrastructure all face similar challenges: high accuracy may coexist with latent instability and modality-specific disagreements. Incorporating temporal and trust-focused metrics ensures responsible operation and reduces the risk of silent failures [6, 14].

By emphasizing TDI and complementary trust metrics, our approach promotes operational transparency and supports dynamic decision-making across domains where human operators or automated controllers rely on continuous predictions.

7 Conclusion

This work presents a comprehensive, deployment-focused evaluation of trust degradation in multimodal time-series predictive maintenance systems operating on edge hardware. Unlike conventional assessments that rely solely on classification accuracy, our analysis demonstrates that real-world reliability cannot be fully captured without considering temporal stability, confidence drift, inter-modality disagreement, and composite metrics such as the Trust Degradation Index (TDI). Although the system maintained consistently high accuracy across temperature, vibration, and acoustic modalities, the longitudinal analyses revealed latent reliability challenges that emerge during continuous operation. Notably, cumulative confidence drift and weighted inter-modality disagreement highlighted periods of silent degradation where fused predictions appeared stable while underlying modalities diverged, underscoring the limitations of conventional performance metrics.

The introduction of the TDI metric enabled a quantitative, time-series view of trust evolution, integrating multiple dimensions of predictive reliability into a single interpretable measure. Deployment-focused plots of temporal stability, cumulative confidence drift, and modality disagreement revealed that trust degradation is neither uniform nor immediately observable, emphasizing the need for continuous monitoring and risk-aware maintenance scheduling. These findings extend beyond predictive maintenance, suggesting similar vulnerabilities in other time-critical, edge-deployed AI systems, including industrial IoT, healthcare monitoring, and smart infrastructure applications.

Future research should focus on several directions:

(1) **Extended deployment studies:** Evaluating trust dynamics over weeks and months across diverse industrial environments (mining, manufacturing, energy generation) to assess long-term drift patterns, seasonal effects, and sensor aging impacts on reliability metrics.

(2) **Adaptive mitigation strategies:** Developing online learning algorithms and dynamic confidence recalibration methods that use TDI feedback to trigger automated model updates, sensor recalibration protocols, or dynamic modality weighting adjustments.

(3) **Real-time monitoring dashboards:** Integrating TDI into operator interfaces with configurable alert thresholds, historical trend visualization, and interpretable explanations linking trust degradation to specific operational events.

(4) **Human-in-the-loop validation:** Conducting controlled studies with maintenance operators to quantify how TDI-informed decisions improve maintenance timing accuracy, reduce false alarms, and enhance overall system trust in operational settings.

(5) **Cross-domain generalization:** Applying trust metrics to autonomous vehicles, medical diagnostic devices, and smart grid systems to establish TDI as a general framework for edge AI reliability assessment beyond predictive maintenance.

Overall, this study underscores the importance of shifting evaluation practices from static, offline accuracy metrics to dynamic, trust-aware methodologies for edge-deployed, multimodal predictive maintenance systems. By explicitly quantifying and visualizing the evolution of predictive trust, practitioners can make more informed deployment decisions, improve operational safety, and enhance confidence in AI-assisted maintenance applications.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion* 76 (2021), 243–297. doi:10.1016/j.inffus.2021.05.008

[2] Stalin Arciniegas, Dulce Rivero, Jefferson Piñan, Elizabeth Diaz, and Francklin Rivas. 2025. IoT Device for Detecting Abnormal Vibrations in Motors Using TinyML. *Discover Internet of Things* 5, 1 (2025), 41. doi:10.1007/s43926-025-00142-4

[3] Firas Bayram and Bestoun S. Ahmed. 2025. Towards Trustworthy Machine Learning in Production: An Overview of the Robustness in MLOps Approach. *Comput. Surveys* 57, 5 (2025), 1–35. doi:10.1145/3708497

[4] Jovani Dalzochio, Rafael Kunst, Edison Pignaton, Alecio Binotto, Sandip Sanyal, Jose Favilla, and Jorge Barbosa. 2020. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Computers in Industry* 123 (December 2020), 103298. doi:10.1016/j.compind.2020.103298

[5] Raúl de la Fuente, Luciano Radrigan, and Anibal S. Morales. 2025. Enhancing Predictive Maintenance in Mining Mobile Machinery through a Hierarchical Inference Network. *IEEE Access* PP (2025). doi:10.1109/ACCESS.2025.3557405 Early Access.

[6] Kiavash Fathi, Tobias Theodor Kleinert, and Hans Wernher van de Venn. 2024. Trustworthy Machine Learning Operations for Predictive Maintenance Solutions. In *Proceedings of the 8th European Conference of the Prognostics and Health Management Society*, Vol. 8. 1039–1042. doi:10.36001/phme.2024.v8i1.3966

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 1321–1330.

[8] Andrew K. S. Jardine, Daming Lin, and Dragan Banjevic. 2006. A Review on Machinery Diagnostics and Prognostics. *Mechanical Systems and Signal Processing* 20, 7 (2006), 1483–1510. doi:10.1016/j.ymssp.2005.09.012

[9] Jay Lee, Behrad Bagheri, and Chao Jin. 2014. Introduction to Cyber Manufacturing. *Manufacturing Letters* 2, 1 (2014), 1–5. doi:10.1016/j.mfglet.2014.01.002

[10] Yaguo Lei, Bin Yang, Xinwei Jiang, Feng Jia, Naipeng Li, and Asoke K. Nandi. 2020. Applications of Machine Learning to Machine Fault Diagnosis: A Review and Roadmap. *Mechanical Systems and Signal Processing* 138 (2020), 106587. doi:10.1016/j.ymssp.2019.106587

[11] Aleksandra Nastoska, Bojana Jancheska, Maryan Rizinski, and Dimitar Trajanov. 2025. Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries. *Electronics* 14, 13 (2025), 2717. doi:10.3390/electronics14132717

[12] Selcuk Selcuk. 2016. Predictive Maintenance, Its Implementation and Latest Trends. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 231, 9 (2016), 1670–1679. doi:10.1177/0954405415601640

[13] Oscar Serradilla, Ekhi Zugasti, Jon Rodriguez, and Urko Zurutuza. 2022. Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Applied Intelligence* 52 (2022), 10934–10964. doi:10.1007/s10489-021-03004-y

[14] Junwei Su and Shan Wu. 2025. Temporal-Aware Evaluation and Learning for Temporal Graph Neural Networks. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI '25, Vol. 39)*. AAAI Press. doi:10.1609/aaai.v39i19.34273

[15] Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X. Gao. 2019. Deep Learning and Its Applications to Machine Health Monitoring. *Mechanical Systems and Signal Processing* 115 (2019), 213–237. doi:10.1016/j.ymssp.2018.05.050