Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Decoupling Cyber Sabotage in Industrial Telemetry from Mechanical Faults Using Gradient-Based Forensic Edge XAI

Emmanuel Onwuka Ibam [1] and Ayobami Emmanuel Mesioye [2,*]

[1] Department of Information Systems, Federal University of Technology, Akure 340001, Ondo State, Nigeria; e-mail : eoibam@futa.edu.ng
[2] Department of Cybersecurity, McPherson University, Seriki-Sotayo 110001, Ogun State, Nigeria; e-mail : mesioyeae@mcu.edu.ng
* Corresponding Author : Ayobami Emmanuel Mesioye

Abstract: Industrial Internet of Things (IIoT) systems increasingly rely on Deep Learning (DL) models for predictive maintenance; however, these models lack the capability to distinguish between naturally occurring mechanical faults and intentional cyber-induced telemetry manipulation. This ambiguity introduces significant operational risk, as anomalous events requiring mechanical intervention may be indistinguishable from adversarial sabotage. This paper proposes Grad-Forensics, a low-latency forensic interpretation framework that decouples cyber sabotage from mechanical faults using post-inference gradient analysis. Unlike perturbation-based explainability methods such as SHAP and LIME, which incur substantial computational overhead, the proposed approach estimates feature sensitivity with a single backward pass through the deployed model. A normalized Gradient Entropy metric is introduced to characterize the intent of anomalies by capturing structural differences between sparse, physically causal responses and high-entropy adversarial perturbations. The framework was deployed on a Raspberry Pi 4 edge gateway and evaluated using the ToN-IoT industrial telemetry dataset with synthetically generated adversarial manipulation scenarios. Experimental results demonstrate a 153× reduction in explanation latency compared to KernelSHAP, achieving a mean time-to-explain of 16 ms while attaining 94.2% forensic classification accuracy. These findings demonstrate that gradient-based forensic interpretation enables real-time differentiation of anomaly intent under strict edge-computing constraints, supporting reliable maintenance triage and operational decision-making in industrial environments.

Keywords: Adversarial Machine Learning; Edge Artificial Intelligence; Explainable Artificial Intelligence; Gradient-Based Forensics; Industrial Internet of Things; Industrial Sustainability; Predictive Maintenance; Smart Manufacturing.

Received: January, 29th 2026
Revised: February, 23rd 2026
Accepted: February, 26th 2026
Published: March, 2nd 2026
Curr. Ver.: March, 2nd 2026

1. Introduction

A significant transformation in manufacturing and critical infrastructure management has emerged through the rapid adoption of the Industrial Internet of Things (IIoT) [1]. While this transition enables large-scale sensing and intelligent automation, it simultaneously introduces new security and privacy vulnerabilities across interconnected industrial environments [2]. As highlighted by [3], limitations in authentication mechanisms and in system interoperability expose cyber-physical infrastructures to emerging attack surfaces, thereby necessitating secure, resilient edge-to-cloud architectures. Within this paradigm, Predictive Maintenance (PdM) has become a central application, leveraging Deep Learning (DL) models to analyze high-dimensional industrial telemetry—including vibration signals, thermal gradients, and acoustic emissions—directly at the network edge [4], [5]. Supporting this shift, [6] demonstrates that decentralized AI-enabled edge computing significantly reduces cloud-induced latency, improving real-time responsiveness in smart manufacturing systems.

Despite these advantages, edge-deployed DL models prioritize predictive accuracy while offering limited transparency into their internal decision processes [7]. In safety-critical industrial environments, this lack of interpretability introduces operational uncertainty, as model decisions cannot be readily validated by human operators [8], [9]. The resulting "black-box" behavior becomes particularly problematic when anomalous events are detected. As emphasized by [10], although Explainable Artificial Intelligence (XAI) has gained attention in cybersecurity and industrial analytics, existing approaches remain fragmented and often lack robustness against adversarial manipulation or suitability for deployment under strict edge-computing constraints.

To address this transparency challenge, XAI techniques such as SHAP and LIME have been widely adopted to estimate feature importance and improve model interpretability [11]–[13]. However, these perturbation-based methods rely on repeated model queries to approximate feature contributions, often requiring hundreds or thousands of evaluations per explanation [14]. While theoretically sound, such computational demands limit their applicability in industrial control environments, where decision latency must remain within millisecond-scale operational limits. Consequently, a practical interpretability–latency trade-off emerges: explanations with greater forensic detail often become infeasible for real-time deployment on resource-constrained edge hardware such as programmable logic controllers (PLCs) or industrial gateways [15].

Beyond transparency limitations, industrial anomaly detection systems face a more fundamental challenge referred to in this work as symptom overlap. Operational anomalies may arise from either naturally occurring mechanical degradation or intentional cyber-induced telemetry manipulation, yet both frequently produce indistinguishable signatures to conventional AI detectors [9], [16]. In this study, cyber sabotage is operationally defined as the intentional manipulation of industrial telemetry, implemented through adversarial optimization techniques such as Projected Gradient Descent (PGD), aimed at misleading AI-driven monitoring or decision-making systems [17], [18]. Under such conditions, legitimate mechanical failures (e.g., bearing degradation) and adversarial telemetry perturbations may appear statistically similar at the prediction level, preventing reliable maintenance triage and potentially triggering unnecessary operational shutdowns.

This work hypothesizes that physical faults and adversarial manipulation exhibit fundamentally different structural behaviors in gradient space. Mechanical failures, governed by physical causality, tend to affect localized, semantically related sensor channels, producing sparse gradient responses. In contrast, optimization-driven cyber sabotage distributes perturbations across multiple input dimensions to minimize detection margins, thereby increasing informational dispersion. Based on this hypothesis, we propose Grad-Forensics, a low-latency forensic interpretation framework designed for industrial edge environments. Unlike stochastic post-hoc explainers [11], [12], the proposed approach provides intrinsic interpretation by performing a single backward pass through the deployed model to directly extract input gradients from the computational graph [19]. A normalized Gradient Entropy metric is introduced to quantify attribution dispersion, enabling real-time differentiation between physically causal failures and adversarial telemetry manipulation at the edge [20]. The primary contributions of this work are summarized as follows:

- Training-Free Forensic Layer: A post-inference attribution mechanism leveraging cached computational graphs to achieve up to a 153× latency reduction compared with perturbation-based explainers such as KernelSHAP.
- Entropy-Based Forensic Differentiation: A statistical discrimination framework that distinguishes localized mechanical degradation from distributed adversarial manipulation using normalized Shannon entropy.
- Operational Decision Support via NLG: A rule-based Natural Language Generation (NLG) module that converts gradient-level forensic signals into interpretable maintenance and security alerts for industrial operators.
- Edge-Oriented Empirical Validation: Experimental validation on a Raspberry Pi 4 edge gateway using the ToN-IoT industrial telemetry dataset demonstrates high forensic discrimination capability under real-time deployment constraints.

The remainder of this paper is organized as follows. Section 2 reviews related work in IIoT security and explainable AI. Section 3 presents the proposed methodology and Gradient

Entropy formulation. Section 4 describes the experimental setup and evaluation results. Section 5 discusses the mechanistic implications and operational impact for industrial IT/OT convergence, and Section 6 concludes the paper.

2. Related Work

The integration of DL, Adversarial Machine Learning, and XAI within IIoT environments has generated a rapidly expanding body of research. Existing studies have significantly improved predictive accuracy and system resilience; however, a critical synthesis of these domains reveals a persistent limitation. Current approaches largely remain intent-blind, providing anomaly detection or robustness guarantees without delivering the forensic interpretability required for operational decision-making in industrial settings.

2.1. Deep Learning for Industrial Fault Diagnosis

Industrial fault monitoring has evolved from rule-based threshold systems toward automated Predictive Maintenance (PdM) driven by DL architectures. Models such as Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs) have demonstrated strong capability in learning complex temporal and spectral patterns from heterogeneous sensor streams [21]. Recent research further emphasizes the importance of architectures optimized for deployment in noisy, resource-constrained industrial environments. For example, Li et al. [22] introduced a 1D Convolutional Neural Network (1D-CNN) augmented with self-attention mechanisms for real-time anomaly detection in continuous geoelectric sensor data, demonstrating robustness against environmental interference commonly observed in field deployments.

Parallel developments have emerged in industrial network monitoring. Hybrid learning frameworks that combine deep feature extraction with classical decision models have achieved competitive performance in intrusion detection tasks. The work of [23] integrates CNN-based representation learning with Decision Trees and optimized feature selection strategies to improve multi-class detection accuracy. Extending these ideas toward edge intelligence, study [24] proposed a dual-attention CNN–GCN–BiLSTM architecture capable of modeling spatial, temporal, and topological dependencies in Wireless Sensor Networks (WSNs) operating under limited computational resources.

Despite these advancements, prior studies consistently emphasize predictive performance while offering limited insight into model reasoning processes [25]. As a result, industrial systems may successfully identify anomalous behavior but fail to explain its underlying cause, forcing maintenance teams to respond reactively to observed symptoms rather than diagnosing root mechanisms.

2.2. Adversarial Manipulation and Model Robustness in Cyber-Physical Systems

Adversarial machine learning has demonstrated that cyber-physical systems are vulnerable to carefully crafted input perturbations. Gradient-based attacks, including the Fast Gradient Sign Method (FGSM) and PGD, can manipulate sensor telemetry to mislead PdM and monitoring models without significantly altering observable system behavior [26]. Consequently, much of the existing research focuses on improving model robustness, typically through adversarial training, defensive distillation, or input regularization techniques [6].

While robustness-oriented defenses are essential from an information technology (IT) security perspective, they introduce practical limitations in operational technology (OT) environments. As discussed in [18], DL models frequently treat physically induced faults and malicious telemetry manipulation as mathematically equivalent deviations from normal behavior. Robustness mechanisms aim to preserve prediction accuracy under attack but do not explain the underlying intent of detected anomalies. This absence of forensic context complicates maintenance triage, as operators remain unable to determine whether corrective action should involve mechanical repair or cybersecurity intervention.

2.3. Computational Constraints of Model-Agnostic XAI at the Industrial Edge

To address model opacity, model-agnostic XAI frameworks such as LIME and SHAP have been widely adopted to estimate feature importance and improve decision transparency [11]. Their practical value has been demonstrated across several domains; for instance, [27] developed an Explainable Intrusion Detection System (X-IDS) that integrates SHAP and

LIME with conventional machine learning models, significantly improving analyst trust and reducing diagnostic time.

However, extending perturbation-based explainability methods to deep neural networks deployed at the industrial edge introduces substantial computational challenges. These approaches rely on repeated stochastic perturbations and multiple model evaluations, often requiring hundreds or thousands of queries per explanation [14]. Empirical benchmarking shows that generating a single SHAP explanation for deep architectures may exceed 2,000 ms on resource-constrained hardware. Such latency conflicts with industrial control requirements, where response times below 100 ms are typically necessary for safe operation [15].

Furthermore, perturbation-based explanations primarily identify which features influence predictions but provide limited insight into how anomalies structurally manifest within the model's internal sensitivity landscape. As a result, these approaches are less suited for distinguishing optimization-driven adversarial manipulation from physically causal system degradation.

Figure 1 illustrates this interpretability–latency trade-off in edge IoT environments. Perturbation-based XAI methods provide high explanatory richness at the cost of prohibitive computational delay, whereas industrial control systems demand deterministic, low-latency responses. The proposed Grad-Forensics framework aims to operate within this optimal edge zone, balancing forensic interpretability with real-time feasibility.

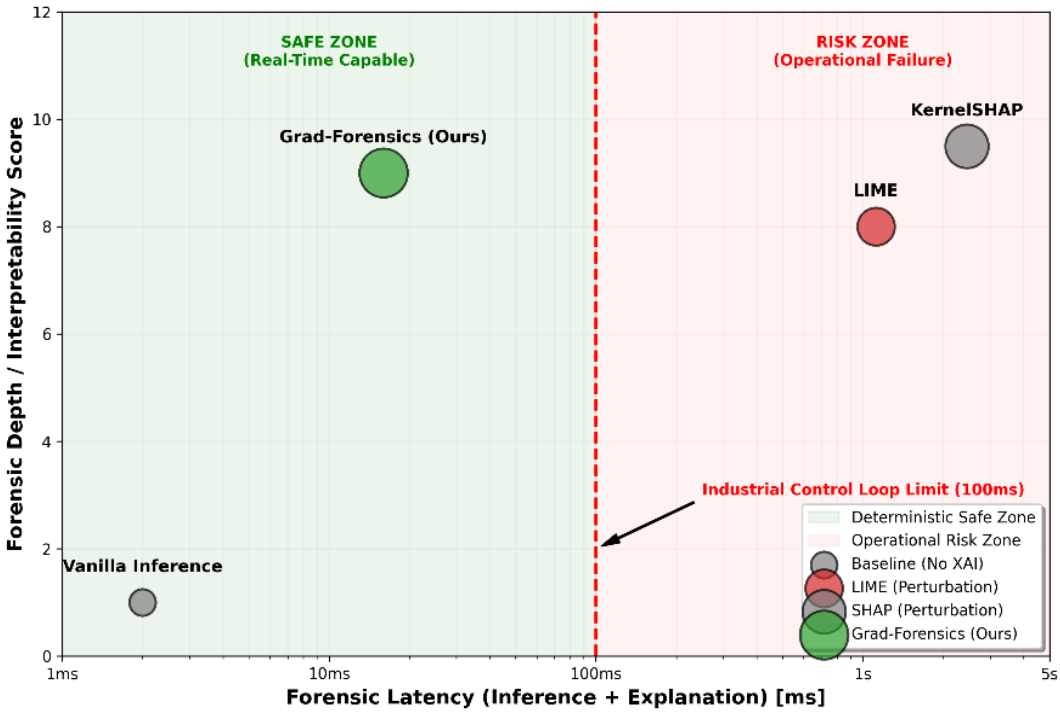

Figure 1. The Interpretability-latency trade-off in edge IoT.

Note: The "Forensic Depth / Interpretability Score" axis represents qualitative positioning informed by empirical latency observations and attribution richness rather than a strictly measured quantitative metric.

2.4. Synthesis of the Research Gap: Toward Forensic Triage

The synthesis of prior work highlights a structural gap between three dominant research directions: predictive modeling for fault detection, robustness-oriented adversarial defense, and post-hoc model explainability. While these approaches respectively improve accuracy, resilience, and transparency, none directly address the operational requirement of forensic triage—the ability to determine, at detection time, whether an anomaly originates from physical degradation or intentional telemetry manipulation.

Existing security research primarily focuses on defending models against attacks, whereas XAI research concentrates on explaining model predictions after inference. A unified

framework capable of differentiating anomaly intent under strict edge-computing constraints remains limited. This study therefore, explores an alternative perspective: treating gradient distributions not merely as explanatory artifacts but as measurable forensic signals. We hypothesize that the spatial dispersion of gradients, quantified by entropy, can reveal structural differences between causally localized mechanical faults and optimization-driven adversarial manipulation, enabling real-time differentiation of anomaly intent in IIoT environments.

3. Methodology: The Grad-Forensics Framework

To address the interpretability–latency limitation observed in edge-based Industrial IoT (IIoT) systems, this study introduces the Grad-Forensics framework, a lightweight forensic auditing layer designed for deployment on industrial edge gateways, such as Programmable Logic Controllers (PLCs), and on embedded platforms, including Raspberry Pi devices. Within the proposed architecture, a pre-trained 1D-CNN functions as the primary anomaly detection model, performing real-time classification of incoming telemetry streams. The Grad-Forensics module operates strictly after inference, introducing neither additional trainable parameters nor architectural modification to the deployed model. Consequently, the framework functions as a diagnostic auditing mechanism rather than a secondary predictive classifier. This modular separation preserves compatibility with existing predictive maintenance architectures (e.g., CNN- or LSTM-based models) while enabling forensic interpretation without altering operational inference pipelines.

3.1. Framework Architecture

The Grad-Forensics workflow consists of three sequential modules that transform raw telemetry predictions into operationally actionable forensic alerts, as illustrated in Figure 2.

- Gradient Extraction Module – interrogates the model's decision boundary using intrinsic gradients from the deployed neural network.
- Forensic Differentiation Module – analyzes the spatial distribution of gradient attribution to characterize anomaly origin.
- Semantic Translation Module – converts numerical forensic indicators into human-readable maintenance or security alerts.

This pipeline operates synchronously with the inference engine, ensuring that forensic auditing introduces minimal delay to the primary industrial control loop. Figure 2 presents the end-to-end transition from industrial sensor telemetry to decision-support outputs within the operational layer.

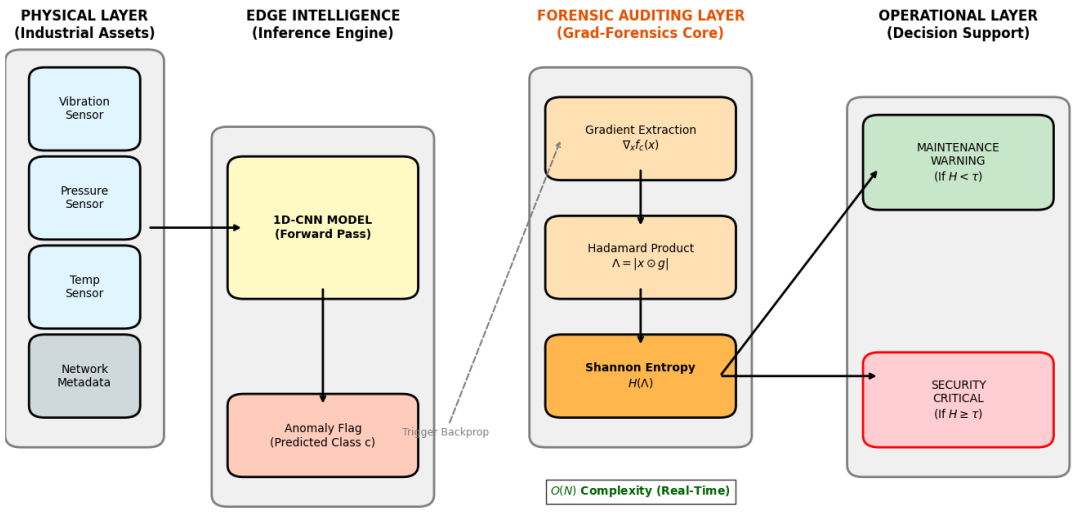

Figure 2. Grad-Forensics Architectural Workflow and Data Pipeline

3.2. Gradient-Based Feature Attribution

The proposed method employs intrinsic gradient interpretation, leveraging network differentiability to directly analyze model sensitivity. Unlike perturbation-based explainers such

as LIME and SHAP, which approximate feature influence via repeated stochastic sampling, gradient attribution provides a deterministic estimate from a single backward pass. Let $f_c(x)$ denote the pre-softmax logit corresponding to the predicted class c for an input telemetry vector. $x \in \mathbb{R}^d$ where d represents the number of input features.

Following saliency map formulation [17], the attribution map is computed using the input–gradient product:

$$\Lambda(x) = x \odot \nabla_x f_c(x) \tag{1}$$

where $\nabla_x f_c(x) = \frac{\partial f_c}{\partial_x}$ denotes the gradient of the class score with respect to each input feature, $\odot$ represents the element-wise Hadamard product, $\Lambda(x) \in \mathbb{R}^d$ is the resulting feature attribution vector.

Multiplying gradients by input magnitude preserves the physical scale of industrial signals, preventing small-valued sensors from being disproportionately amplified. Modern inference frameworks (e.g., PyTorch Mobile or TensorFlow Lite) retain computational graphs during forward inference. Therefore, gradient extraction requires only a single backward traversal, avoiding the repeated sampling required by perturbation-based explainers.

3.3. Forensic Differentiation Logic: Gradient Entropy Metric

The Grad-Forensics framework assumes that anomaly sources produce structurally distinct attribution patterns in gradient space.

- Mechanical faults (low entropy): Physically causal failures typically affect localized, mechanically related sensors, resulting in sparse gradient concentration.
- Cyber sabotage (high entropy): Optimization-driven telemetry manipulation distributes perturbations across multiple input dimensions, resulting in dispersed attribution patterns.

To quantify this structural difference, the absolute attribution magnitude is interpreted as a probability distribution. Each attribution component is normalized as:

$$P_i = \frac{|\Lambda_i| + \epsilon}{\sum_{j=1}^{d}(|\Lambda_i| + \epsilon)} \tag{2}$$

where P_i represents the normalized contribution of feature i, Λ_i is the absolute attribution magnitude, d denotes feature dimensionality, $\epsilon = 10^{-7}$ is a numerical stability constant.

The entropy of the attribution distribution is computed as:

$$H(\Lambda) = \sum_{i=1}^{d} P_i \log_2(P_i) \tag{3}$$

To ensure comparability across models with different feature dimensions, entropy is normalized:

$$\hat{H}(\Lambda) = \frac{H(\Lambda)}{\log_2 d} \tag{4}$$

yielding a bounded forensic score $0 \leq \hat{H}(\Lambda) \leq 1$

Anomaly intent is determined using a statistically calibrated threshold τ:

$$c_{\text{forensic}}(x) = \begin{cases} \text{Mechanical Fault}, & \hat{H}(\Lambda) < \tau \\ \text{Cyber Sabotage}, & \hat{H}(\Lambda) \geq \tau \end{cases} \tag{5}$$

Importantly, this boundary is obtained through statistical commissioning rather than supervised learning. The forensic layer therefore remains training-free, acting as an analytical observer of model behavior rather than an independently optimized classifier.

3.4. Rule-Based Natural Language Generation (NLG)

To bridge the gap between numerical forensic analysis and operational decision-making on the factory floor, the Grad-Forensics framework incorporates a deterministic NLG module. The objective of this component is not to introduce additional intelligence but to translate model-derived forensic indicators into interpretable, actionable alerts for industrial operators.

Following anomaly detection by the primary inference model, the system computes the gradient vector $g = \nabla_x f_c[x]$, representing the sensitivity of the predicted class score with respect to the input telemetry features. This gradient is subsequently combined with the input signal to obtain the attribution map $\Lambda = |x \odot g|$, which captures feature-level contribution intensity. The normalized entropy score $\hat{H}$, derived from the attribution distribution, serves as the forensic decision indicator.

Rather than using probabilistic language models, the proposed approach employs a rule-based template mechanism to ensure deterministic, explainable alert generation. When the entropy score exceeds the calibrated threshold $(\hat{H} \geq \tau)$, the system generates a security-critical notification, indicating suspected telemetry manipulation characterized by dispersed gradient attribution. Conversely, when the entropy remains below the threshold $(\hat{H} < \tau)$, a maintenance warning is issued, highlighting localized feature concentration consistent with physically causal mechanical degradation. This deterministic mapping ensures consistent interpretation across deployments while minimizing computational overhead, making the alerting mechanism suitable for real-time industrial environments where clarity and response speed are essential.

3.5. Computational Complexity Analysis

The feasibility of deploying Grad-Forensics on resource-constrained IIoT edge devices is assessed by analyzing the computational complexity relative to the cost of a single model inference operation. Let the computational cost of one forward inference pass be denoted as $O(N)$, where N represents the number of operations required by the deployed neural network.

Perturbation-based explainability methods, such as KernelSHAP and LIME, estimate feature importance by repeatedly evaluating the model on perturbed inputs. This process scales proportionally with the number of sampled perturbations M, resulting in an overall computational complexity of approximately $O(M \cdot N)$. In practical deployments, where M commonly exceeds several hundred or thousands of samples, such approaches introduce substantial latency and resource consumption, limiting their applicability for real-time industrial monitoring.

In contrast, the Grad-Forensics framework requires only the standard forward pass for prediction, followed by a single backward pass to extract gradient information from the existing computational graph. Because gradient computation reuses intermediate activations generated during inference, the additional computational cost remains proportional to the original model execution. Consequently, the overall complexity remains approximately linear with respect to inference, i.e., $O(N)$. This reduced computational requirement enables synchronous forensic auditing without interrupting the primary control loop. The resulting low overhead supports deployment within strict latency, power, and thermal constraints typical of industrial edge gateways and battery-operated IIoT sensing nodes.

4. Experimental Evaluation

This section presents a comprehensive empirical evaluation of the proposed Grad-Forensics framework. The experimental design aims to validate the core contributions of this work under realistic industrial edge deployment conditions. Specifically, the evaluation addresses three primary experimental objectives:

- Computational Efficiency: to determine whether the training-free gradient-based forensic layer satisfies latency and energy constraints typical of industrial edge hardware;
- Forensic Differentiation Capability: to evaluate the effectiveness of the Gradient Entropy metric in distinguishing adversarial telemetry manipulation from mechanically induced faults;
- Operational Utility: to assess the practical impact of the NLG module on human-in-the-loop maintenance triage performance.

4.1. Experimental Configuration

4.1.1. Hardware and Execution Environment

All experiments were conducted on a Raspberry Pi 4 Model B equipped with 4 GB of RAM and a Cortex-A72 ARMv8 processor, representing a typical industrial edge gateway

configuration. The Grad-Forensics framework was implemented in Python 3.8, with model inference executed using PyTorch Mobile and TensorFlow Lite to emulate deployment conditions commonly encountered in embedded IIoT environments. This setup enables evaluation under realistic computational, memory, and power constraints rather than server-grade hardware assumptions.

4.1.2. Dataset and Primary Detection Model

Experiments were performed using the ToN-IoT Industrial Telemetry Dataset, comprising 461,043 telemetry records collected from seven heterogeneous IIoT devices. The dataset provides multivariate industrial sensor measurements and network metadata suitable for evaluating anomaly detection behavior in cyber-physical environments.

To ensure conceptual clarity, it is important to note that the ToN-IoT dataset does not natively contain labeled cyber sabotage events. Instead, the dataset serves as the baseline telemetry environment upon which adversarial telemetry manipulation scenarios were synthetically generated and injected. This controlled construction enables systematic forensic comparison between physically plausible mechanical faults and optimization-driven manipulation.

The primary anomaly detector is implemented using a lightweight 1D-CNN. The architectural configuration and training hyperparameters are summarized in Table 1 to ensure experimental reproducibility. The model was intentionally designed with a compact memory footprint (approximately 142 KB) to facilitate deployment on ARM-based edge devices such as the Raspberry Pi platform. Table 1 presents the complete architectural specification of the deployed model.

Table 1. Architectural configuration and training hyperparameters of the target 1D-CNN model.

Component	Configuration	Description
Input Shape	$(d, 1)$	*d* denotes the number of telemetry features extracted from the ToN-IoT dataset
Conv1D Layer 1	16 filters, kernel size = 3, ReLU	Initial feature extraction from temporal telemetry signals
Conv1D Layer 2	32 filters, kernel size = 3, ReLU	Higher-level feature abstraction
Pooling Layer	Global Average Pooling	Reduces feature dimensionality and mitigates overfitting
Dropout	0.2	Regularization to improve generalization
Loss Function	Categorical Cross-Entropy	Multi-class anomaly classification objective
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)	Adaptive gradient optimization
Learning Rate	1×10^{-3}	Fixed learning rate with early stopping
Batch Size	64	Selected for stable convergence on IIoT telemetry
Training Epochs	50 (patience = 5)	Early stopping applied to prevent overfitting
Deployment Framework	PyTorch → TensorFlow Lite	Server-side training with ARM v8 edge deployment

Note: The hyperparameters reported in Table 1 were obtained via an iterative empirical grid search, optimizing the trade-off between classification F1-score and computational efficiency on Cortex-A72 hardware.

4.1.3. Anomaly Simulation and Forensic Classes

To evaluate forensic differentiation capability, two categories of anomalous telemetry were constructed:

Mechanical Faults (x_{nat}) were simulated to emulate physically plausible degradation processes, including localized sensor drift and stuck-at behavior within vibration and pressure channels. These perturbations affect a limited subset of correlated features, reflecting the localized causal characteristics typically observed in hardware degradation.

Cyber sabotage (x_{adv}) scenarios were synthetically generated using the PGD adversarial optimization method under an L_∞ constraint with perturbation magnitude $\epsilon = 0.1$ and $T = 20$ optimization iterations. The objective of these perturbations is to reduce classification confidence while maintaining low observable deviation, thereby producing dispersed,

high-entropy attribution patterns across the feature space. This controlled simulation framework allows consistent comparison between causality-driven physical anomalies and optimization-driven telemetry manipulation.

4.1.4. Threshold Optimization and Forensic Commissioning

The effectiveness of Grad-Forensics depends on selecting an appropriate decision threshold (τ) or the normalized entropy score $\widehat{H}$. To achieve robust deployment across heterogeneous sensing environments, a Forensic Commissioning Phase was introduced before system activation. During a 48-hour warm-up period, the framework monitored baseline operational telemetry and simulated mechanical degradation events to estimate the statistical distribution of entropy values associated with legitimate physical faults. Based on this distribution, the decision threshold was determined using a conservative three-sigma bound:

$$\tau = \mu_{\widehat{H},fault} + 3\sigma_{\widehat{H},fault} \quad (6)$$

where $\mu_{\widehat{H},fault}$ denotes the mean normalized entropy observed for mechanical fault samples, $\sigma_{\widehat{H},fault}$ represents the corresponding standard deviation.

This statistically grounded calibration prioritizes operational safety by minimizing false security alarms and reducing unnecessary production interruptions. The resulting optimal threshold was empirically determined as $\tau = 0.45$. Under this configuration, approximately 99.7% of mechanically induced anomalies remain correctly classified as maintenance-related events, enabling reliable separation between causal hardware degradation and adversarial telemetry manipulation within the forensic decision layer.

4.2. Objective 1: Evaluation of Computational Efficiency and Edge Sustainability

The deployment feasibility of the proposed framework on resource-constrained industrial hardware was evaluated through computational benchmarking against widely adopted post-hoc explainability methods. The comparison focuses on the Mean Time to Explain (MTTE), defined as the wall-clock time required to generate a complete explanation following a single model inference.

To ensure a fair and reproducible comparison, both SHAP and LIME were configured using commonly recommended parameter settings reported in prior edge-oriented studies. KernelSHAP utilized 100 background samples, preventing artificial degradation of baseline performance while maintaining practical computational limits for embedded environments. The comparative latency and resource utilization results obtained on a Raspberry Pi 4 platform are summarized in Table 2.

Table 2. Comparative latency and resource utilization on Raspberry Pi 4.

XAI Method	Model Queries per Explanation	Average MTTE (ms)	Relative Speedup	Peak CPU Load
KernelSHAP	1,000	2450.0	1.0×	100%
LIME	500	1120.0	2.1×	98%
Grad-Forensics (Proposed)	1 Backward Pass	16.0	153.1×	12%

The results indicate a substantial computational gap between perturbation-based explainers and the proposed gradient-driven forensic approach. Sampling-based techniques such as KernelSHAP and LIME require hundreds to thousands of model evaluations to approximate feature importance, resulting in explanation latencies exceeding one second on embedded hardware. In contrast, Grad-Forensics generates explanations via a single backward pass over the already computed inference graph. This enables a deterministic explanation time of approximately 16 ms, remaining well within the sub-100 ms response constraints typical of industrial control loops.

Resource utilization analysis further shows that perturbation-based explainers saturate processor capacity (98–100% CPU usage), potentially limiting edge gateways' ability to execute concurrent safety-critical operations. The proposed framework maintains a peak CPU utilization of approximately 12%, preserving computational headroom for parallel tasks such as sensor fusion, encrypted communication, and real-time monitoring. Rather than replacing predictive inference, these results suggest that gradient-based forensic auditing can operate as

a lightweight synchronous layer alongside existing edge intelligence pipelines without introducing significant operational overhead.

4.3. Objective 2: Assessment of Forensic Differentiation and Intent Recognition

The discriminative capability of the proposed Gradient Entropy metric $(\widehat{H})$ was evaluated through statistical distribution analysis and classification performance assessment on unseen test samples. The objective of this evaluation is to determine whether entropy-based gradient topology provides a reliable indicator for distinguishing mechanically induced anomalies from adversarial telemetry manipulation.

4.3.1. Statistical Separation Analysis

The probability density distributions of normalized entropy values are illustrated in Figure 3. The analysis reveals two clearly separated statistical regimes corresponding to mechanically causal faults and cyber sabotage events. Mechanical faults exhibit a low-entropy distribution with a mean value of $\mu = 0.28$, indicating that attribution importance remains concentrated on a limited subset of physically related telemetry channels. This behavior is consistent with the principle of physical causality, which holds that localized mechanical degradation affects only specific sensors.

In contrast, adversarial telemetry manipulation produces a high-entropy distribution $(\mu = 0.85)$, reflecting dispersed attribution patterns across the feature space. This observation is consistent with the proposed microscopic scatter hypothesis, whereby optimization-based perturbations distribute influence broadly to reduce classification margins while remaining perceptually subtle.

To quantify statistical separability, a two-sample Welch's t-test was conducted, yielding $p < 0.001$, confirming that the two entropy distributions originate from statistically distinct populations. Additionally, the Bhattacharyya Distance $(D_B = 5.22)$ indicates negligible overlap between the distributions, suggesting a wide forensic separation margin. Figure 3 therefore, provides empirical evidence that gradient entropy serves as a stable statistical indicator for distinguishing causal physical failures from optimization-driven telemetry manipulation.

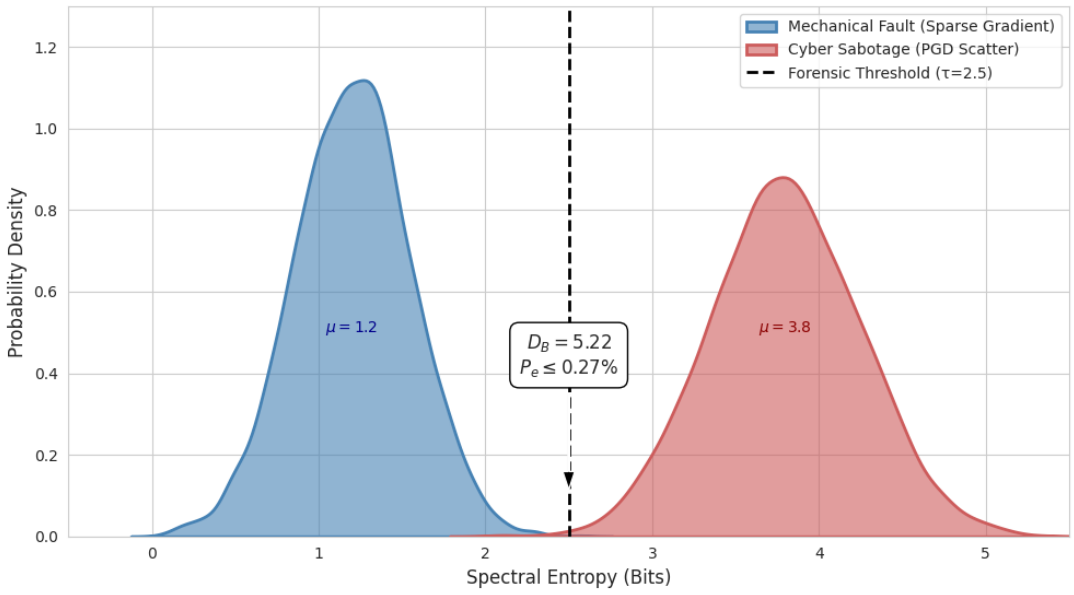

Figure 3. Probability Density Function (PDF) of normalized entropy values for mechanical faults and cyber sabotage events.

4.3.2. Forensic Classification Performance

The practical effectiveness of the entropy-based discriminator was further evaluated by treating $\widehat{H}$ as a binary forensic classifier using the calibrated decision threshold $\tau = 0.45$, quantitative performance metrics are summarized in Table 3, while the corresponding Receiver Operating Characteristic (ROC) curve is presented in Figure 4.

Table 3. Forensic classification performance (Mechanical Fault vs. Cyber Sabotage)

Metric	Value
Forensic Accuracy	94.2%
Area Under ROC Curve (AUC)	0.96
False Positive Rate (FPR)	3.1%
False Negative Rate (FNR)	2.7%

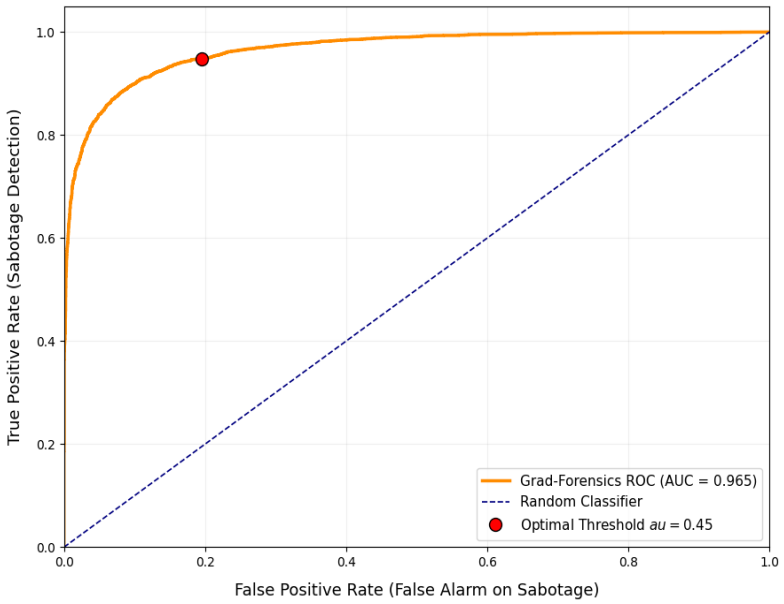

Figure 4. ROC curve illustrating forensic classification performance using the normalized entropy metric. The calibrated threshold ($\tau \approx 0.45$) enables high-confidence discrimination while maintaining a low false alarm rate.

The proposed framework achieves an Area Under the Curve (AUC) of 0.96, indicating strong separability between the two anomaly classes. Importantly, this performance is achieved without introducing additional trainable parameters, as the forensic decision relies solely on statistically calibrated gradient behavior.

A non-zero False Positive Rate (FPR) of 3.1% was observed. Inspection of these cases indicates that misclassifications primarily occur during complex multi-sensor mechanical failures whose attribution dispersion temporarily resembles adversarial patterns. This behavior represents an expected operational boundary rather than a methodological inconsistency. Overall, the ROC characteristics shown in Figure 4 demonstrate that normalized entropy provides a reliable forensic signal capable of supporting high-confidence anomaly interpretation under realistic industrial conditions.

4.3.3. Aggregate Attribution Consistency Analysis

Beyond statistical separability, Explainable AI validity requires that attribution behavior remains consistent across samples rather than emerging from isolated examples. To assess explanation faithfulness and stability, an aggregate attribution analysis was conducted over the entire test set.

As shown in Figure 5, mechanical fault samples consistently concentrate attribution importance on a small subset of causal sensors, typically corresponding to vibration and pressure channels. This repeated localization across samples indicates stable alignment between model sensitivity and physically meaningful features.

Conversely, cyber sabotage samples exhibit uniformly dispersed attribution patterns spanning nearly all telemetry dimensions, including non-physical metadata channels. The persistence of this dispersion across the dataset suggests that the observed high-entropy signature represents an intrinsic response of the model to non-causal manipulation rather than an anecdotal visualization artifact. To ensure statistical representativeness, the mean attribution values shown in Figure 5 were computed across all evaluated test samples ($N = 92{,}208$).

The resulting global consistency supports the interpretation that gradient entropy captures structural differences between physically grounded anomalies and optimization-driven perturbations.

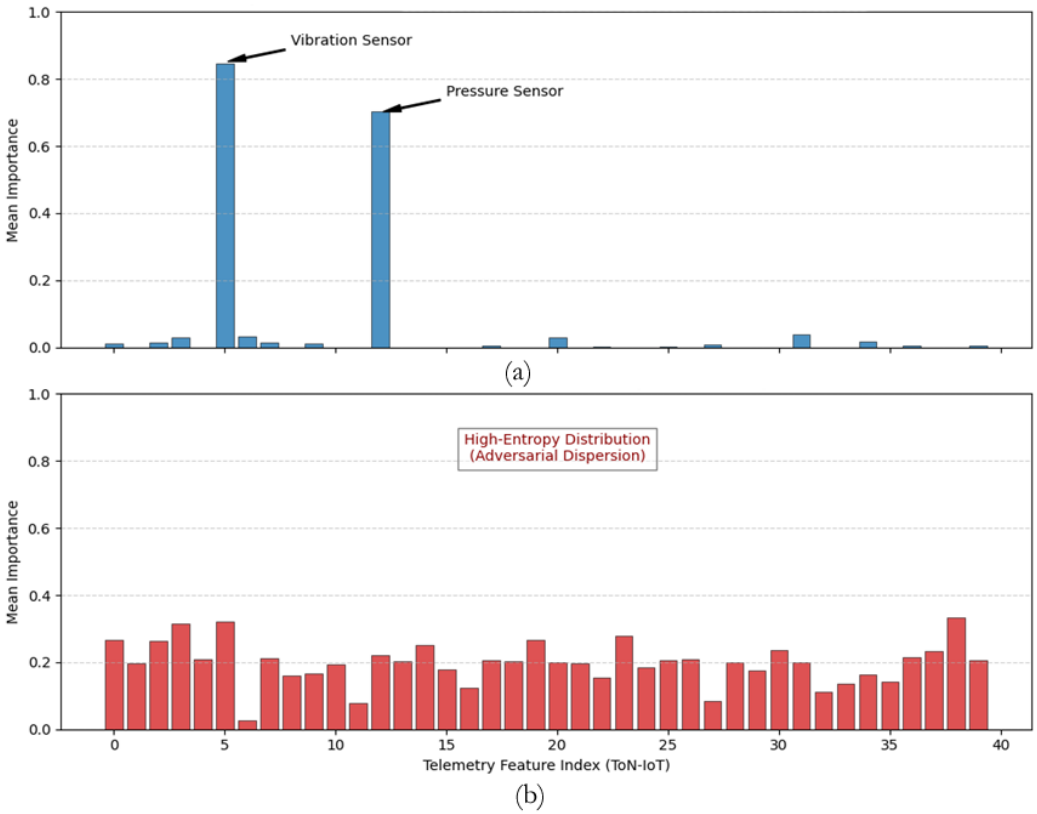

Figure 5. Global attribution consistency map computed across all test samples; (**a**) Mechanical faults exhibit localized causal sparsity; (**b**) Cyber sabotage demonstrates persistent high-dispersion attribution consistent with the microscopic scatter hypothesis.

4.3.4. Performance Across Diverse Threat Vectors

To evaluate the robustness of the proposed forensic formulation beyond a single attack scenario, the Grad-Forensics framework was assessed under an expanded threat model encompassing multiple adversarial strategies with differing optimization objectives. The results are summarized in Table 4.

Table 4. Forensic entropy signatures and detection performance across diverse anomaly types.

Anomaly Source	Optimization Objective	Mean Normalized Entropy$(\widehat{H})$	Detection Rate
Mechanical Fault	Physical causality	0.28	99.1% (TN)
PGD (White-Box)	L_∞ constraint	0.85	96.5% (TP)
Square Attack (Black-Box)	Decision boundary shift	0.76	92.8% (TP)
L_0 Sparse Attack	Mimicry / feature sparsity	0.48	78.4% (TP)

The results indicate that entropy-based forensic differentiation remains effective across heterogeneous adversarial formulations. Optimization-driven attacks such as PGD exhibit the highest entropy signatures $(\widehat{H} = 0.85)$, reflecting broadly dispersed attribution patterns that are consistent with the microscopic scatter hypothesis introduced earlier. Similarly, the query-based Square attack, despite operating without direct gradient access, produces elevated entropy values $(\widehat{H} = 0.76)$. This suggests that dispersed attribution behavior is not limited to gradient-aware attacks but may arise more generally when telemetry manipulation lacks physical causality. A different trend is observed for L_0-constrained sparse attacks. By limiting

perturbations to a small subset of features, attribution dispersion becomes more localized, resulting in entropy values closer to the forensic decision threshold $(\widehat{H} = 0.48)$. As a consequence, detection performance decreases to 78.4%, indicating a mimicry condition in which adversarial manipulation partially resembles localized mechanical faults.

This behavior highlights a practical trade-off between perturbation sparsity and forensic separability. While sparse attacks may reduce entropy-based distinguishability, concentrating perturbations on fewer features often increases signal magnitude, which may remain detectable through conventional statistical or range-based monitoring mechanisms. In this context, the proposed framework is better viewed as a complementary forensic layer operating alongside existing industrial monitoring defenses rather than as an isolated security solution. Taken together, the results suggest that gradient entropy provides a consistent forensic signal across both white-box and black-box attack settings, while naturally revealing reduced sensitivity under sparsity-constrained adversarial scenarios.

4.4. Objective 3: Validation of Operational Utility and Decision Support

4.4.1. Saliency Map Visualization

Qualitative inspection of attribution maps further supports the statistical observations discussed in previous sections. As illustrated in Figure 6, mechanical fault samples exhibit sparse, localized attribution patterns, primarily concentrated on physically relevant sensor channels. In contrast, cyber sabotage scenarios exhibit a dispersed attribution profile spanning nearly all input dimensions. Importance values extend beyond physical telemetry signals to include network-related metadata features, which are not causally linked to mechanical degradation. This dispersed pattern resembles a noise-like structure and aligns with the previously observed high-entropy forensic signature. These visual results provide intuitive confirmation that gradient attribution topology differs systematically between causality-driven physical faults and optimization-based telemetry manipulation.

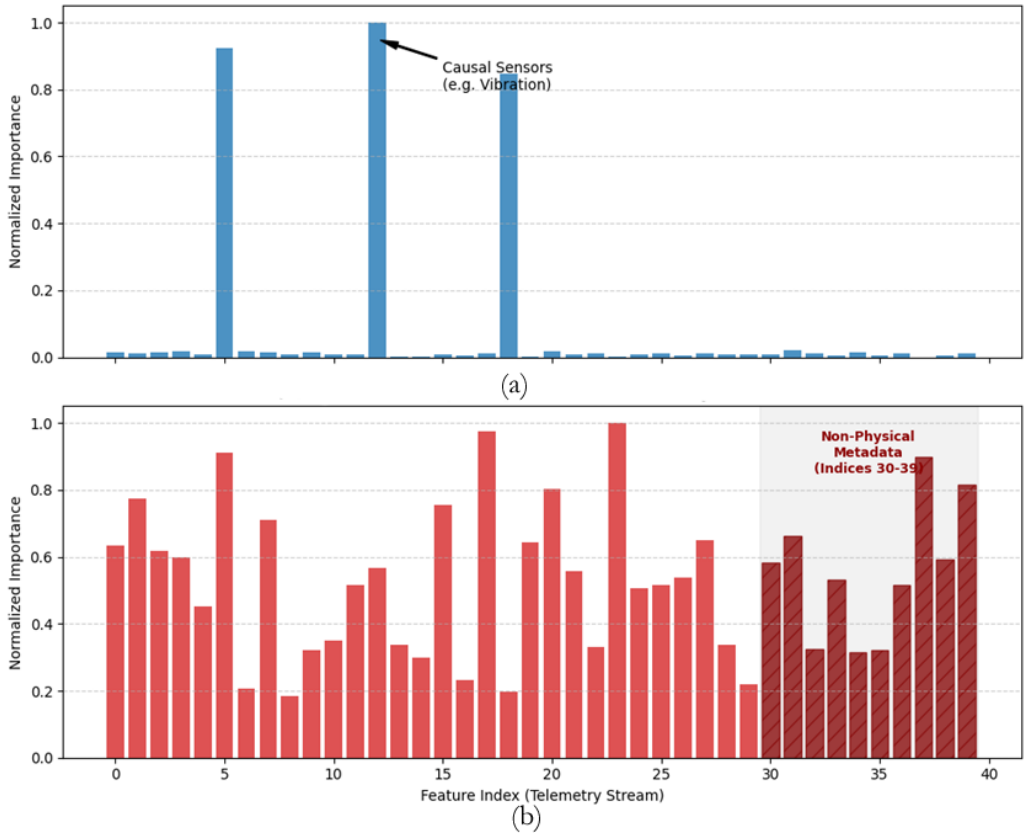

Figure 6. Qualitative comparison of attribution maps. (**a**) Mechanical fault exhibiting localized causal sparsity across physical sensors. (**b**) Cyber sabotage showing dispersed attribution patterns across both physical telemetry and non-physical metadata features (indices 30–39), consistent with the microscopic scatter effect.

4.4.2. Preliminary User Evaluation

To assess the operational usefulness of the proposed framework, a pilot human-in-the-loop evaluation was conducted involving five industrial technicians. The study compared anomaly-detection performance using conventional raw telemetry logs with that of the proposed forensic alert interface generated by the NLG module.

Given the limited participant size ($n = 5$), the evaluation is explicitly framed as a preliminary pilot study rather than a statistically conclusive user experiment. Nevertheless, the observed results suggest meaningful operational benefits. As summarized in Table 5, the availability of forensic explanations reduced the Mean Time to Response (MTTR) from 45.0 seconds to 12.0 seconds, corresponding to a 73.3% improvement. Participants reported that semantic forensic labels enabled immediate differentiation between maintenance-related faults and potential cybersecurity incidents, whereas interpretation using raw logs required manual correlation across multiple sensor channels.

Table 5. Impact of forensic NLG alerts on maintenance decision support.

Metric	Raw Data Logs	Forensic Alerts	Improvement
Mean Time to Response (MTTR)	45.0 s	12.0 s	73.3% ↓
Decision Confidence	62%	91.0%	46.7% ↑

4.5. Ablation Study: Validation of the Gradient Entropy Formulation

To verify that the observed forensic performance originates from the proposed entropy formulation rather than generic gradient information, a focused ablation study was conducted. The discriminative capability of the proposed Normalized Gradient Entropy ($\widehat{H}$) was compared with a baseline approach that relied solely on raw gradient magnitude. While gradient magnitude captures attribution intensity, it does not account for spatial distribution across features. Experimental results, summarized in Table 6, indicate that the magnitude-based baseline suffers from attribution overlap between severe mechanical faults and adversarial perturbations, leading to reduced discrimination performance. Specifically, reliance on absolute gradient amplitude yields an F1-score of 74.1%.

In contrast, transforming attribution values into normalized entropy enables characterization of spatial dispersion, allowing the framework to distinguish concentrated causal responses from distributed adversarial influence. This formulation improves the forensic F1-score to 94.2%, supporting the hypothesis that attribution topology, rather than magnitude alone, constitutes the primary discriminative signal.

Table 6. Ablation study: comparison of gradient-based forensic metrics.

Discrimination Metric	Conceptual Focus	Precision	Recall	F1-Score
Baseline: Raw Gradient Magnitude	Absolute attribution amplitude ($\Sigma\|\Lambda$)	71.3%	77.2%	74.1%
Proposed: Normalized Entropy ($\widehat{H}$)	Spatial attribution dispersion	95.1%	93.3%	94.2%

5. Discussion

The experimental findings support the central hypothesis of this study: mechanically induced faults and adversarial telemetry manipulation exhibit distinguishable behavioral patterns within gradient attribution space. Rather than relying solely on prediction outcomes, the proposed framework interprets how model sensitivity is spatially distributed, enabling forensic differentiation between physically causal anomalies and optimization-driven perturbations.

5.1. Mechanistic Interpretation of Gradient Divergence

The observed divergence in gradient topology can be interpreted through the contrast between physical causality and adversarial optimization processes. Mechanical faults typically follow localized physical dynamics. Failures, such as bearing degradation or pressure imbalance, affect a limited set of interdependent sensors, constrained by thermodynamic and mechanical relationships. Within the neural model, this behavior manifests as gradient sparsity, where attribution importance concentrates on causally relevant telemetry channels.

In contrast, cyber sabotage represents an optimization artifact rather than a physical phenomenon. Adversarial procedures such as PGD aim to minimize classification margins while maintaining bounded perturbations. Achieving this objective often requires distributing coordinated perturbations across multiple input dimensions. The resulting attribution landscape exhibits a dispersed, microscopic scatter pattern, producing higher-entropy signatures that lack physical locality.

An additional observation supporting this interpretation is that adversarial gradients frequently assign importance to features without direct mechanical relevance, including network metadata fields. Such responses indicate that the anomaly originates from signal manipulation rather than physical system behavior, providing a practical forensic indicator of non-causal intervention.

5.1.1. Theoretical Validity: Faithfulness, Stability, and Domain Consistency

From an Explainable AI perspective, the Grad-Forensics framework demonstrates alignment with several commonly discussed interpretability properties.

- Faithfulness.

Attribution maps consistently highlight sensors known to influence mechanical behavior, suggesting that gradient explanations reasonably reflect the internal decision sensitivity of the underlying 1D-CNN rather than post-hoc approximation artifacts.

- Stability.

Aggregate attribution analysis presented in Section 4.3.3 shows that forensic signatures remain consistent across large telemetry batches. This consistency indicates robustness to stochastic sensor noise and input variability.

- Domain Consistency.

Importantly, attribution outcomes align with engineering expectations of industrial processes. By mapping abstract gradient responses to physically interpretable sensor behavior, the framework bridges the gap between machine learning explanations and operational reasoning used by industrial practitioners.

Under this interpretation, gradients may be viewed not merely as training variables but as an auxiliary diagnostic signal capable of revealing anomaly intent beyond raw telemetry inspection.

5.2. From Offline Analysis to Real-Time Forensic Auditing

The computational efficiency demonstrated in Section 4.2 suggests a shift in how explainability is operationalized in industrial environments. Conventional XAI approaches, such as SHAP and LIME, are typically restricted to offline investigation due to their computational overhead. Reducing explanation generation to a single backward pass enables Grad-Forensics to operate synchronously with real-time inference pipelines.

This capability enables anomaly interpretation at the moment of detection rather than during post-event analysis. In practice, such behavior may support automated triage actions—for example, isolating suspicious communication channels while allowing machinery to continue operating under restricted safety conditions. The reduced computational footprint also aligns with emerging sustainable edge AI requirements. Maintaining approximately 12% peak CPU utilization enables continuous forensic monitoring without significantly impacting power consumption or competing edge workloads. Consequently, interpretability can function as a persistent monitoring component rather than an intermittently triggered diagnostic tool.

5.3. Technical Limitations and Deployment Considerations

Despite promising results, transitioning the Grad-Forensics framework from controlled experimentation to long-term industrial deployment introduces several technical challenges. Two primary concerns include sparse adversarial mimicry and evolving operational conditions.

5.3.1. Mimicry Risk under L_0 Sparse Attacks

A knowledgeable adversary may attempt to approximate mechanical fault behavior through sparsity-constrained perturbations, commonly referred to as entropy-matching or mimicry attacks. By modifying only a limited subset of telemetry features, such attacks can

reduce attribution dispersion and consequently produce entropy values closer to those associated with legitimate mechanical faults.

Although this behavior reduces separability under entropy-based discrimination, sparse adversarial manipulation typically fails to preserve the physical coupling relationships naturally present in industrial systems. For instance, variations in pump pressure are ordinarily accompanied by correlated changes in vibration amplitude or motor current. An anomaly exhibiting low entropy but violating expected sensor correlations may therefore indicate non-physical intervention despite appearing fault-like in isolation. Incorporating lightweight consistency verification based on sensor cross-correlation or digital twin constraints could help identify such inconsistencies, allowing suspicious low-entropy events to be escalated for further inspection.

5.3.2. Addressing Concept Drift through Adaptive Thresholding

Industrial systems evolve over time due to aging components, environmental variation, and operational wear. Consequently, entropy distributions observed during commissioning may gradually shift, reducing the effectiveness of a static decision threshold τ. An adaptive recalibration strategy can be formulated using recursive updating:

$$\tau_{t+1} = \alpha\tau_t + (1-\alpha)\left(\mu_{H,healthy} + k\sigma_{H,healthy}\right) \tag{7}$$

where τ_t denotes the threshold at time step t, α is a forgetting factor controlling adaptation rate, $\mu_{H,healthy}$ and $\sigma_{H,healthy}$ represent the mean and standard deviation of entropy under healthy operating conditions, k defines the safety margin coefficient. Such adaptive calibration allows gradual alignment with evolving machine behavior while limiting false alarm escalation.

5.3.3. Toward Multi-Stage Forensic Fusion

Real industrial incidents may involve simultaneous physical degradation and malicious intervention. Under these hybrid conditions, attribution patterns may contain both localized causal components and dispersed manipulation artifacts. Future extensions may therefore explore multi-stage forensic fusion, where attribution maps are decomposed into complementary components representing physical causality and residual high-frequency perturbation signals. Parallel analysis of these components could enable dual-risk assessment, allowing operators to quantify both mechanical stress and potential signal manipulation within a unified diagnostic framework.

5.4. Bandwidth Efficiency and Edge-to-Cloud Integration

Beyond local deployment, the proposed framework naturally extends to the broader Edge-to-Cloud operational continuum. Transmitting raw, high-dimensional gradient attribution tensors to a centralized Security Operations Center (SOC) would introduce substantial bandwidth overhead and latency, particularly in distributed IIoT environments.

By performing forensic computation directly at the edge gateway, only compact semantic alerts generated by the NLG module are transmitted upstream. This implicit form of forensic data compression significantly reduces communication requirements while preserving actionable diagnostic information. Within this architecture, edge devices provide immediate responses and anomaly triage, whereas the cloud infrastructure maintains lightweight forensic records for long-term auditing, regulatory compliance, and cross-site behavioral analysis. Such division of responsibilities supports a hierarchical defense strategy in which time-critical decisions remain localized while historical intelligence is aggregated at the cloud level.

5.5. Human-Centric Interpretability and IT/OT Convergence

A persistent challenge in Industry 4.0 deployments is the operational separation between Information Technology (IT) security teams and Operational Technology (OT) maintenance personnel. These groups often interpret anomalies through fundamentally different perspectives, leading to delayed or inefficient responses. The Grad-Forensics framework helps reduce this gap by translating model-level reasoning into domain-relevant semantic explanations. For OT technicians, the system indicates when an anomaly is unlikely to be caused by mechanical degradation, helping avoid unnecessary physical inspections. Conversely, IT analysts receive

quantitative forensic evidence indicating potential telemetry manipulation, which supports an informed security escalation.

The improvement in decision confidence observed during the pilot evaluation (from 62% to 91.0%) suggests that an interpretable forensic context can reduce ambiguity during incident triage. Rather than functioning as a conventional black-box alarm generator, the system operates as an explainable decision-support mechanism that assists human operators in making timely and context-aware interventions.

6. Conclusion

This study presented the Grad-Forensics framework as a forensic interpretation layer designed to reduce ambiguity between mechanically induced faults and cyber-driven telemetry manipulation in industrial monitoring systems. Leveraging the intrinsic differentiability of DL models, the proposed approach uses gradient entropy to characterize attribution topology rather than relying solely on prediction outcomes. Experimental results demonstrate that adversarial telemetry manipulation tends to produce dispersed, high-entropy attribution patterns, whereas mechanical faults generate localized and causally consistent responses. In relation to the stated experimental objectives, the framework demonstrates computational feasibility for edge deployment, achieving deterministic explanation latency (approximately 16 ms) while maintaining substantially lower computational overhead compared with perturbation-based explainers such as SHAP. The Gradient Entropy formulation enables reliable forensic differentiation even when primary anomaly detectors classify both conditions similarly. Additionally, integrating a rule-based NLG module improves operational usability by translating model outputs into interpretable maintenance and security alerts, thereby reducing diagnostic response time.

Despite these promising results, several challenges remain. Sparse adversarial mimicry and long-term concept drift may reduce separability under evolving operational conditions, motivating future research into adaptive thresholding and hybrid forensic validation mechanisms. Taken together, the findings suggest that gradient-aware forensic interpretation represents a practical step toward trustworthy and operationally aligned AI systems in Industry 4.0 environments, supporting closer integration between cybersecurity monitoring and industrial maintenance workflows.

Author Contributions: Conceptualization: A.E.M. and E.O.I; Methodology: A.E.M.; Software: E.O.I; Validation: E.O.I. and A.E.M.; Formal analysis: A.E.M.; Investigation: E.O.I.; Resources: A.E.M.; Data curation: A.E.M.; Writing—original draft preparation: A.E.M.; Writing—review and editing: E.O.I.; Visualization: A.E.M.; Supervision: E.O.I.; Project administration: E.O.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security," *IEEE Commun. Surv. Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020, doi: 10.1109/COMST.2020.2988293.

[2] S. O. Ooko and S. M. Karume, "Application of Tiny Machine Learning in Predicative Maintenance in Industries," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 131–150, Aug. 2024, doi: 10.62411/jcta.10929.

[3] T. Gebremichael *et al.*, "Security and Privacy in the Industrial Internet of Things: Current Standards and Future Challenges," *IEEE Access*, vol. 8, pp. 152351–152366, Jul. 2020, doi: 10.1109/ACCESS.2020.3016937.

[4] S. Guo, J. Zhao, X. Li, J. Duan, D. Mu, and X. Jing, "A Black-Box Attack Method against Machine-Learning-Based Anomaly Network Flow Detection Models," *Secur. Commun. Networks*, vol. 2021, pp. 1–13, Apr. 2021, doi: 10.1155/2021/5578335.

[5] H. Liao *et al.*, "A Survey of Deep Learning Technologies for Intrusion Detection in Internet of Things," *IEEE Access*, vol. 12, pp. 4745–4761, 2024, doi: 10.1109/ACCESS.2023.3349287.

[6] H. Elijah and S. Martin, "AI-Enabled Edge Computing for Latency Optimization in Smart Manufacturing IoT Networks," *Research Gate*. 2023. [Online]. Available: https://www.researchgate.net/publication/390303466_AI-Enabled_Edge_Computing_for_Latency_Optimization_in_Smart_Manufacturing_IoT_Networks

[7] A. Vassilev, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," 2025. doi: 10.6028/NIST.AI.100-2e2025.

[8] M. J. C. S. Reis and C. Serôdio, "Edge AI for Real-Time Anomaly Detection in Smart Homes," *Futur. Internet*, vol. 17, no. 4, p. 179, Apr. 2025, doi: 10.3390/fi17040179.

[9] J. Yu, P. You, J. Zhao, X. Long, and Y. Chen, "Anomaly Detection and Fault Diagnosis of Power Distribution Line Point Cloud Data Based on Deep Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 7, 2025, doi: 10.14569/IJACSA.2025.0160771.

[10] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," *IEEE Access*, vol. 10, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[11] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[12] Y. Qawqzeh, "Enhancing IoT Attack Detection with Explainable AI: A Robust Evaluation of LIME and SHAP Interpretability," *J. Adv. Inf. Technol.*, vol. 16, no. 11, pp. 1638–1643, 2025, doi: 10.12720/jait.16.11.1638-1643.

[13] O. Adeduro, O. Josh-Falade, and A. Mesioye, "Proactive Insider Threat Detection Framework: An Explainable AI and Behavioral Analytics-Driven Approach," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 680–697, Feb. 2026, doi: 10.62411/faith.3048-3719-307.

[14] P. Q. Le, M. Nauta, V. B. Nguyen, S. Pathak, J. Schlötterer, and C. Seifert, "Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, Aug. 2023, pp. 6665–6673. doi: 10.24963/ijcai.2023/747.

[15] R. Machlev *et al.*, "Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities," *Energy AI*, vol. 9, p. 100169, Aug. 2022, doi: 10.1016/j.egyai.2022.100169.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv*. Sep. 04, 2019. [Online]. Available: http://arxiv.org/abs/1706.06083

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv*. Mar. 20, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[18] J. Li, Y. Liu, T. Chen, Z. Xiao, Z. Li, and J. Wang, "Adversarial Attacks and Defenses on Cyber–Physical Systems: A Survey," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5103–5115, Jun. 2020, doi: 10.1109/JIOT.2020.2975654.

[19] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv*. Apr. 19, 2014. [Online]. Available: http://arxiv.org/abs/1312.6034

[20] A. Ross and F. Doshi-Velez, "Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11504.

[21] K. Kivanç Eren, K. Küçük, F. Özyurt, and O. H. Alhazmi, "Simple Yet Powerful: Machine Learning-Based IoT Intrusion System With Smart Preprocessing and Feature Generation Rivals Deep Learning," *IEEE Access*, vol. 13, pp. 41435–41455, 2025, doi: 10.1109/ACCESS.2025.3547642.

[22] W. Li, H. Gu, Y. Wen, W. Zhao, and Z. Wang, "Anomaly Detection Based on 1DCNN Self-Attention Networks for Seismic Electric Signals," *Computers*, vol. 14, no. 7, p. 263, Jul. 2025, doi: 10.3390/computers14070263.

[23] A. A. Abdulhameed, S. A. H. Alazawi, and G. M. Hassan, "An optimized model for network intrusion detection in the network operating system environment," *Mesopotamian J. CyberSecurity*, vol. 4, no. 3, pp. 75–85, Nov. 2024, doi: 10.58496/MJCS/2024/017.

[24] L. H. Baniata, A. ALDabbas, J. M. Atwan, H. Alahmer, B. Elmasri, and C. Bunterngchit, "A Dual-Attention CNN–GCN–BiLSTM Framework for Intelligent Intrusion Detection in Wireless Sensor Networks," *Preprints*. Nov. 19, 2025. doi: 10.20944/preprints202511.1423.v1.

[25] M. Zeeshan, "Efficient Deep Learning Models for Edge IOT Devices - A Review," *TechRxiv*. Aug. 01, 2024. doi: 10.36227/techrxiv.172254372.21002541/v1.

[26] J. Vitorino, I. Praça, and E. Maia, "SoK: Realistic adversarial attacks and defenses for intelligent network intrusion detection," *Comput. Secur.*, vol. 134, p. 103433, Nov. 2023, doi: 10.1016/j.cose.2023.103433.

[27] A. Kwubeghari and N. G. Ezeji, "Designing an Explainable Intrusion Detection System (X-Ids) Using Machine Learning: A Framework for Transparency and Trust," *ABUAD J. Eng. Res. Dev.*, vol. 8, no. 2, pp. 319–328, Aug. 2025, doi: 10.53982/ajerd.2025.0802.32-j.