*future internet*

MDPI

*Article*

# Explainable AI for Securing Perception-Layer Sensor Data in IoT Environmental Danger Detection Systems

**Taha Al-Jadir [1], Iván García-Magariño [2] and Raquel Lacuesta Gilaberte [1,*]**

[1] Systems and Information Engineering, Escuela Universitaria Politécnica de Teruel, University of Zaragoza, c/Atarazana 2, 44003 Teruel, Spain; 829946@unizar.es
[2] Department of Software Engineering of Artificial Intelligence, Instituto de Tecnología del Conocimiento, Complutense University of Madrid, 28040 Madrid, Spain; igarciam@ucm.es
* Correspondence: lacuesta@unizar.es

**Abstract**

This paper presents an explainable defense framework against perception-layer and Man-in-the-Middle (MitM) attacks in Internet of Things (IoT)-based environmental hazard warning systems. These systems rely on heterogeneous sensors (gas, light, sound, temperature, and humidity) whose integrity is crucial for reliable environmental alerts. Perception-layer attacks such as spoofing, jamming, and data injection can compromise sensor readings, while MitM attacks threaten communication reliability. The proposed approach integrates incremental Dynamic Time Warping (DTW) for time-series anomaly detection with a tree-based ensemble classifier (XGBoost), in addition to Shapley Additive Explanations (SHAP) for interpretability. A comparative evaluation framework jointly considers detection performance and explanation quality through metrics including pre-registering a Casual Ground Truth based on network protocol localized Precision@ K feature overlap metrics (Q), instead of relying on subjective human-expert or global rank correlations to quantitively evaluate the explanation transparency. Experimental simulations using an authentic EdgeIIoT-2022 dataset under 3-fold forward–chaining cross-validation demonstrated high detection accuracy and moderated explainability scores. The results prove the framework's ability to detect and explain adversarial behaviors in sensor networks, strengthening trust, transparency, and resilience in safety-critical IoT infrastructures.

**Keywords:** IoT security; environmental detection system; explainable artificial intelligence; man-in-the-middle attack; Distributed Denial of Service; shapely additive explanations; time-series anomaly detection; perception-layer attack



Academic Editors: Paolo Bellavista and Francesco Buccafurri

Received: 30 June 2026
Revised: 16 July 2026
Accepted: 22 July 2026
Published: 24 July 2026



## 1. Introduction

In the field of the Internet of things (IoT), a wide range of security vulnerabilities still pose serious risks to systems relying on distributed sensing [1]. These systems often operate autonomously and continuously, processing large volumes of environmental data for real-time decision-making. Understanding the impact of IoT security leaks has become critical, as several incidents demonstrate the potential severity of such vulnerabilities. Examples include (a) the hacked baby monitors in Ohio and Texas [2], (b) the attacks over devices produced by Acoustic Technology Inc. [3], and (c) the Turning Up the Freeze Distributed Denial of Service (DDoS) attack over environmental control systems [4]. IoT security threats are typically categorized into three major layers: (1) perception-layer attacks, such as botnets, node tampering, jamming, or data spoofing; (2) network-layer attacks, including Man-in-the-Middle (MitM), DDoS, perception-layer attacks that compromise sensor integrity, and

routing manipulation; and (3) application-layer attacks, such as malware or code injection. Among these, perception-layer attacks [5] are particularly concerned in systems that depend heavily on accurate and trustworthy sensor readings for example for considering environmental awareness or user health.

The growing deployment of IoT-based environmental monitoring systems [6] introduces new challenges in security and reliability. These systems, which integrate gas, light, sound, temperature, and humidity sensors, aim to warn users about potential dangers in a location based on the sensed environmental data. The gathered information is processed by embedded microcontrollers and transmitted via Wi-Fi to backend services responsible for real-time alerting through intelligent assistants such as Alexa. Despite their evident usefulness for public safety and environmental monitoring, these systems can become ineffective or even dangerous when perception-layer attacks in this context can modify sensor readings, inject false data, or manipulate the temporal arrival of packets (Inter-Arrival Time), leading to erroneous environmental interpretations. For instance, a malicious alteration in gas concentration measurements [7] could prevent an alert from being triggered during a gas leak, or an artificial light or sound disturbance could simulate nonexistent environmental hazards. Therefore, the robustness of the perception layer becomes essential for maintaining the reliability of the entire warning infrastructure.

This article focuses on perception-layer security in IoT-based environmental hazard warning systems. Specifically, it explores the MitM attack not only as a network routing threat but as a high-fidelity perception-layer intervention targeting sensor data acquisition and proposes defensive mechanisms based on anomaly detection and time-series analysis. By reinforcing the perception layer, the goal is to ensure that alerts generated by these systems remain accurate, timely, and trustworthy, even in the presence of adversarial interference. Moreover, this article explores MitM attacks in network communication when sensors are communicated through Internet to the environmental system. By utilizing ARP-spoofing at the IoT edge, the attacker can subtly warp the temporal signature of sensor streams. We propose a transparent, feature-driven approach using DTW and (XAI) to provide casual evidence of temporal anomalies which lead to detecting these deviations. Unlike black-box detection methods, our approach derives a transparent causal link between the physical network intervention and the resulting anomaly. This ensures the defensive mechanism is not only accurate but methodologically sounds and explainable to security auditors. We do not claim a new intrusion classifier in general; instead, we contribute a reproducible evaluation design for explainable perception-layer data protection in which the explanation metrics are explicit, auditable and aligned with the top ranked features that human analysts inspect.

## 2. Related Work

XAI has been previously applied in cybersecurity in the context of deep learning and other artificial intelligence (AI) approaches. Pawlicki et al. [8] reviewed major XAI techniques and their benefits, systematically mapped the field to identify trends, and highlighted future research directions for integrating explainability into AI-driven cybersecurity systems. However, they did not provide an approach that avoided both perception-layer and MitM attacks for ensuring appropriate functioning in environmental systems with different environmental sensors.

### *2.1. XAI in AI-Driven Cybersecurity Systems*

IoT enables the interconnection of heterogeneous devices capable of sensing, processing, and communicating data to provide context-aware services. These systems have evolved to support critical applications such as environmental monitoring, smart homes, healthcare, and industrial safety. In the case of environmental hazard warning systems, the perception

layer plays a fundamental role, as it is responsible for acquiring reliable measurements from the physical environment through various sensors, including gas, light, sound, temperature, and humidity sensors. For instance, the study of Kong et al. [9] highlights the potential for model transparency; there remains a critical research gap in applying these techniques to detect temporal distortions at the perception layer. Using characteristic index gases, a dynamic discriminant model achieved high accuracy. Comparative analyses confirmed superior performance, offering valuable insights for efficient fire prevention in mining areas. However, this work did not suggest novel XAI cybersecurity mechanisms for ensuring the proper functioning of these environmental systems. In addition, comprehensive surveys, such as systematics mapped by Rjoub et al. [10], have primary XAI techniques (Shapely Additive explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and integrated gradients) and an outlined strategic direction for embedding explainability into autonomous defense systems. Similarly, Capuano et al. [11] provided a structured taxonomy of XAI applications in cybersecurity, emphasizes the need for model-agnostic explanations that can generalize across diverse threat detection scenarios.

Pawlicki et al. [8] further advanced this by mapping future research directions for XAI in deep learning and AI used in cybersecurity, identifying critical gaps in quantitative validation frameworks and the integration of XAI with real-time threat response mechanisms. Even with these conceptual taxonomies, the practical implementation remains limited. A huge part of the existing XAI framework operates as a post hoc diagnostic tool for network packet classification instead of an active component capable of guiding defenses against perception-layer data manipulation and MitM attacks.

Moreover, few studies quantitively leverage XAI alignment (evaluating explanation stability via rank correlation metrics) to validate if the model's local decision aligns consistently with known physical attack vectors across diverse environmental sensors. Miller et al. [12] mentioned this limitation by arguing that most XAI research relies on instinct instead of the practically grounded theories of human explanation, calling for multi-specialized methods to integrate insights from cognitive psychology and social sciences into XAI future design.

### *2.2. Perception-Layer Vulnerabilities and Data Integrity Defenses*

The perception layer represents the entry point for all environmental data, making it the first target for adversaries seeking to compromise the integrity of the system. The attacks at this layer generally aim to manipulate sensor inputs or weaken their functionality to cause incorrect evaluations of environmental conditions. Common examples may include sensor spoofing, where attackers generate artificial triggers to mislead the system. For example, emitting specific gas concentration to falsify readings; signal jamming, which prevents sensors from accurately transmitting data; and data injection attacks, where faked sensor values are introduced into the communication channel. Kong et al. [9] designed an intelligent early warning system for preventing environmental pollution from coal spontaneous combustion using characterized index gases and a dynamic discriminate model to achieve high accuracy in fire detection. While achieving high classification performance, their architecture neglected defensive cybersecurity mechanisms, leaving the system exposed to malicious injections.

In terms of environmental data collection systems, perception-layer attacks can have severe consequences. For instance, spoofing or tampering with gas sensors could prevent the detection of toxic leaks, while manipulating sound or light sensors might trigger false alarms or conceal actual environmental hazards. These vulnerabilities are particularly critical in safety-critical systems designed to warn people about potential dangers in their surroundings. Nasralla et al. [5] proposed a technique for defending against perception-layer attacks in the context of IoT smart furniture for impaired people. They used Dynamic Time

Warping (DTW) comparison for identifying strange daily series compared to the usual daily series as an indicator of a malfunctioning sensor or perception-layer attack. Our current work also uses DTW but adds XAI concepts for making environmental administrators more aware of the reasons behind an attack warning. Moreover, a recent study has started to address this gap by proposing learning-based and data-driven approaches to detect perception-layer anomalies. For example, time-series analysis methods such as DTW and Long Short-Term Memory (LSTM) networks have been applied to identify deviations from normal sensor behavior in smart environments. For instance, Yolaçan and Zaim [13] proposed an LSTM-based Dynamic Compound Weight Mechanism (DCWM) to detect cyberattacks on robotic arms, including replay and subscriber flood attacks. These techniques rely on the temporal correlation and continuity of sensor data to detect inconsistencies that may indicate attacks. In addition, sensor fusion techniques have been explored to cross-validate measurements from multiple sensor types, improving resilience against single-sensor compromise. Nevertheless, Yolaçan and Zaim did not include the application of XAI as the current work does for making environmental systems trustable in the eyes of users.

### *2.3. Multi-Class IoT Intrusion Detection and Communication Security*

Research on IoT security has traditionally focused on network and application layers, where threats such as Man-in the-Middle (MitM) attacks, DDoS, and malware have been extensively analyzed. For instance, Cherian and Varma [14] addressed IoT security challenges, particularly Distributed Denial of Service (DDoS) and MitM attacks, by proposing a Belief-Based Secure Correlation methodology. Integrating IoT with Software-Defined Networking (SDN) and Redstone cryptographic encryption, the framework ensured secure data transmission, dynamic route selection, and effective prevention of DDoS, MitM, and related data attacks. However, the perception layer remains comparatively underexplored, even though it represents the foundation of trust in IoT architectures. Without securing this layer, even the most sophisticated encryption or network defenses become ineffective, as decisions are based on compromised or inaccurate sensor data. The current work provides a joint approach that considers both communication attacks such as MitM and perception-layer attacks in the diversity of IoT sensors in environmental systems. Similarly, Alani et al. [15] and Khedr et al. [16] have advanced MitM detection using the ARP spoofing detection system and muti-classifier ensemble method on the EdgeIIoT-2022 dataset, achieving accuracy exceeding 99.89% with false alarm rates below 0.02%. However, these approaches prioritize network metadata over the underlying time-series telemetry of physical sensors. Concurrently, many studies focus on developing multi-class classification algorithms especially designed for IoT network traffic using benchmark datasets like Edge-IIoT-2022. Ferrage et al. [17] introduced this comprehensive dataset capturing authentic traffic from large-scale IoT testbeds, including protocols such as MQTT, HTTP, and DNS, with fourteen attack categories spanning DoS/DDoS, information gathering, MitM, injection and malware attacks. While these metrics achieved near-perfect metrics (F1-score approaching 0.99 for specific network vectors like MitM), they rarely bridge the gap between network-layer metadata and underlying time-series telemetry of physical sensors.

### *2.4. Emerging Trends: Large Model Agents in IoT Security*

As IoT ecosystems scale into autonomous, decentralized environments, the integration of large model agents (LMAs) and the internet of agents (IoA) have emerged as an active research frontier; recent research provides critical insights into these paradigms:

State-of-the-art cooperation paradigms: Wang et al. [18] comprehensively surveyed the state of the art in large model-based agents, detailing their capacity for multi-agent reasoning, autonomously orchestrating in complex environments, comprehending intent,

and their work maps' ability to cooperate with paradigms including data cooperation, computation cooperation, and knowledge cooperation across cloud-edge-end architecture, showing the potential for LMAs to perform coordinated threat modeling and real-time defense adaptation.

Internet agents (fundamentals and challenges): Wang et al. [19] provided a systematic review of IoA fundamentals, applications and challenges, emphasizing the need for scalable and secure coordination among heterogenous agents; their work identifies key architectural components including capability notification, dynamic discovery, and trust-based task orchestration, which are essential for deploying autonomous defense mechanisms in distributed IoT environments.

Security, Privacy, and Challenges: Wang et al. [20] offered a comprehensive examination of the security and privacy landscape in IoA systems, focusing on four critical aspects (identity authentication threats, cross-agent trust issues, embodied security, and privacy risks); their taxonomy reveals that LMAs introduce unique threat surfaces including prompt injection, hallucination-driven policy failure, and privacy leakages over edge networks (specifically embodied security threats such as sensor spoofing and cross-modal backdoors) which can lead to harmful physical behaviors when agents control IoT infrastructure.

While LMAs offer unparalleled potential for high-level collaboration and future autonomous remediation, their heavy computational footprint limits direct execution on resource-constrained perception-layer microcontrollers. This reinforces the necessity of the proposed framework, which provides lightweight, deterministic, tree-based ensemble models backed by quantitative XAI to ensure edge data before it is ingested by upstream autonomous agents.

### *2.5. Research Gap Analysis*

Despite significant progress across these individual domains, a critical research gap persists at the intersection of temporal feature engineering, multi-class communication security, and quantitative XAI verification; existing frameworks force a trade-off between the mathematical robustness of temporal alignment (for example, DTW) and the systemic visibility offered by explainable model outputs. Table 1 shows the capabilities of recent representative state-of-the-art approaches against our proposed multi-layered defense architecture.

**Table 1.** Comparative analysis for the state of the art (✗: not included, ✓: included).

| Study | Perception-Layer Defense | Network/MitM Protection | Temporal Feature Extraction | Multi-Class Security | XAI Core Integration | Quantitative XAI Verification |
|---|---|---|---|---|---|---|
| Kong et al. (2025) [9] | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Rjoub et al. (2023) [10] | ✗ | ✗ | ✗ | ✗ | ✓ (Theoretical) | ✗ |
| Capuano et al. (2022) [11] | ✗ | ✗ | ✗ | ✗ | ✓ (Theoretical) | ✗ |
| Pawlicki et al. (2024) [8] | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| Miller (2019) [12] | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| Nasralla et al. (2020) [5] | ✓ | ✗ | ✓ (DTW) | ✗ | ✗ | ✗ |
| Yolaçan & Zaim (2025) [13] | ✗ | ✗ | ✓ (LSTM) | ✗ | ✗ | ✗ |
| Cherian & Varma (2022) [14] | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| Alani et al. (2023) [15] | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| Khedr et al. (2023) [16] | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ |
| Wang et al. (2025) [18] | ✗ | ✗ | ✗ | ✗ | ✓ (LLM-based) | ✗ |
| Proposed Framework | ✓ | ✓ | ✓ (DTW + Time-Series) | ✓ | ✓ (SHAP) | ✓ (localized Q) |

## 3. Proposed Approach

### *3.1. Overview*

This work proposes a technique that aims to defend IoT-based environmental danger warning systems against perception-layer attacks and MitM attacks that compromise sensor readings and lead to unreliable or misleading environmental alerts. It also considers MitM attacks in the connection between sensors and the environmental system as well as the network link to the cloud backend. Since the perception layer is responsible for capturing physical-world data through sensors, it represents the most vulnerable point in the IoT stack. Malicious alterations to this layer can undermine higher-layer trust, making it essential to implement intelligent and interpretable defense mechanisms.

The defense model combines time-series anomaly detection with (XAI) achieves two objectives:

- **High Attack Recall:** Identify irregular or adversarial behaviors in sensor readings caused by spoofing, jamming, or data injection.
- **Transparency and Trust:** Provide interpretable explanations for the detected anomalies, ensuring that both developers and end-users understand the rationale behind each detection decision.

This technique is designed for environmental system architectures that include environmental sensors (gas, light, sound, temperature, humidity) connected to a microcontroller for local processing and data transmission to a cloud-based backend. The backend executes the proposed learning-based defense module and triggers alerts through an Alexa-based notification interface when a real environmental hazard is confirmed. Figure 1 shows this architecture of the defense mechanism, in which the normal alert messages (NAMs) include some explanations based on the XAI output.

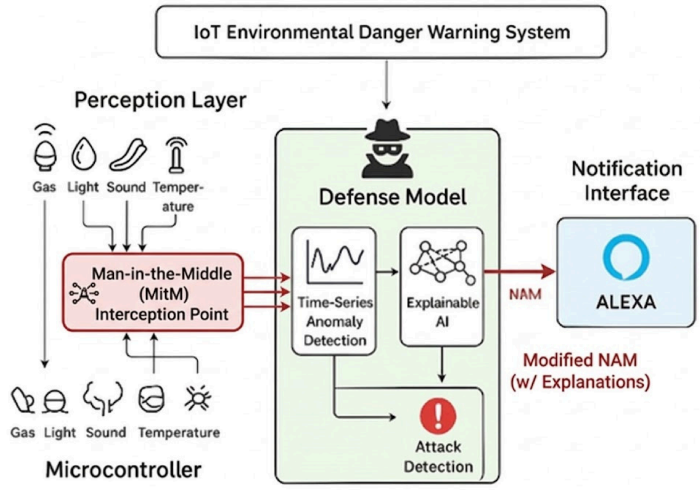


**Figure 1.** High-level layout of the IoT environmental danger warning system. (Note: Raw environmental sensor measurements are Min–Max normalized into unitless numerical features scaled between [0,1] up on network packetization, ensuring consistent features scaling prior to ingestion by the defense model).

The proposed framework consists of a serialized four-stage pipeline designed to ingest, process, evaluate and validate streaming network data in real time. The comprehensive workflow transitions systematically from the initial data ingestion to localized model validation with precision@ k feature overlap as illustrated in Figure 2:

1. **Dynamic stream initialization:** Establishes the real-time context by ingesting streaming live network packets and dynamically extracting the vector baselines for the next tasks.

2. **Incremental DWT time-series extraction:** Uses a sliding window protocol configured to a sequence length of 10 packets to compute dynamic temporal costs.
3. **Multi-objective detection:** Passes the computed temporal metrics to ensemble classifier (XGBoost) to execute highly accurate traffic classification coupled with a SHAP explanation layer for instant predictive attribution.
4. **Localized precision@ k XAI validation**: Extracts the top-k SHAP feature importances and cross-references them against the pre-registered expert casual matrix.

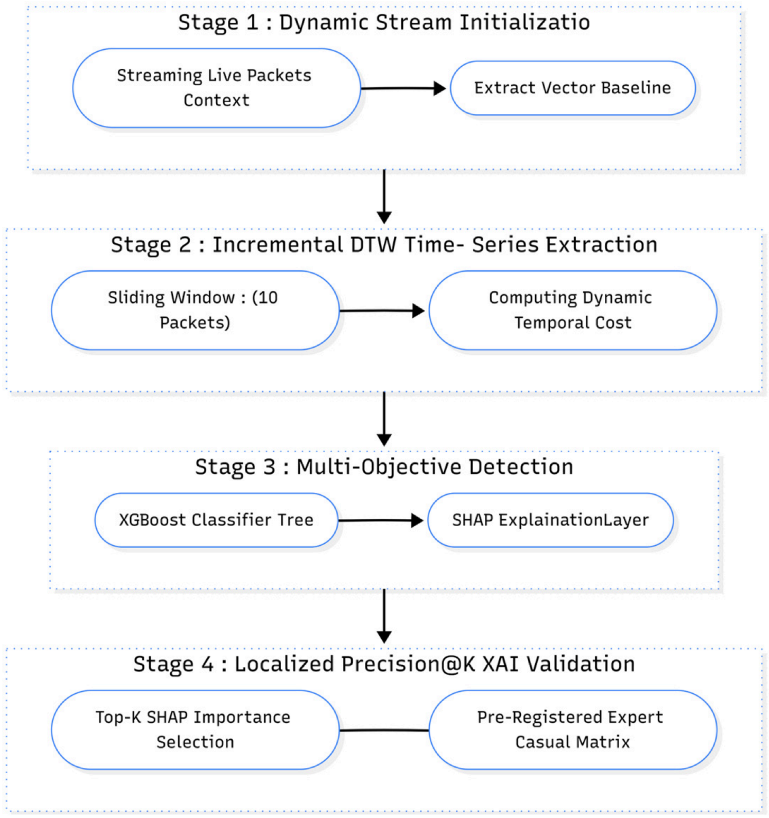


**Figure 2.** Proposed defense methodology diagram.

### *3.2. Learning-Based Detection Module*

The proposed defense mechanism uses a hybrid learning model that integrates distance metrics (Dynamic Time Warping (DTW) or Euclidean distance) for temporal similarity analysis and tree-based ensemble classifiers (e.g., XGBoost or Random Forest); for attack classification, the model constantly reads incoming network traffic, then it implements a dual-routing strategy:

- **Temporal attacks (MitM and similar attacks):** Attackers intercept and change network data, which causes a noticeable "rhythm disruption" in how and when packets arrive. These attacks are routed through our sequential feature engineering pipeline (length and timing) to detect the shift.
  We calculate DTW cost by comparing a sliding window of the current 10 income packets against a pre-registered "Normal" reference window.
  This converts the raw network signature into a "Warping Cost" feature that measures the temporal incongruence. Large deviations in DTW distance values indicate temporal inconsistencies.
- **Non-temporal attacks (SQL Injection and other attacks):** These stealthy attacks do not disrupt network packet schedules or sequence timing; therefore, they bypass the

sequential distance calculation and route directly to the multi-dimensional tree structure of the XGBoost engine, which catches them using the standard packed field vector.

- **Anomaly Classification**: Extracted time-series features (mean, variance, autocorrelation, spectral energy, DTW Cost) blend with standard network features to generate a unified and multi-dimensional feature vector denoted as ($X_{packet}$), used as input to a supervised learning classifier trained to distinguish between normal and attacked sensor behavior.

$$X_{packet} = \left[ \mu \, , \, \sigma^2, \, R_{xx}( \, \tau \, ), \, E_{spectral}, \, Distance - cost \, \right]$$

where μ is the mean, $\sigma^2$ is the variance, $R_{xx}( \, \tau \, )$ is the lag-based autocorrelation, $E_{spectral}$ is the FFT-based spectral energy, and $Distance - cost$ represents the extracted (DTW or Euclidean values).

Figure 3 shows an illustrative example of incongruence detected in a series, in which the attacker switched the sensor wires connecting sensors and processors in a simulated environment.

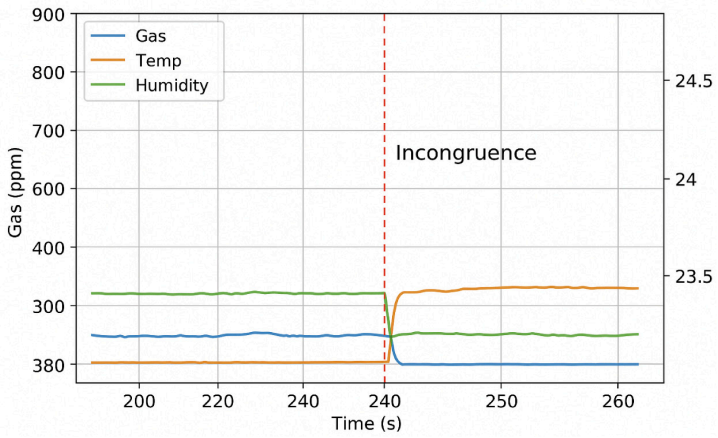


**Figure 3.** Illustrative example of clean incongruence.

To explain the differences to the time series, we used the decomposition of different time series as several series of influential factors that could represent either normal or hazard situations. The explanations are these decompositions of influential factors in each of the parameters. When a hazard is detected, an explanation is provided. When an anomaly is detected, it provides the time series and the comparison with the most similar series to show the large differences. This module provides an estimation of which sensor(s) may be hacked according to the incongruencies.

- **Temporal Split Protocol**: To satisfy the operational constraints and prevent data leakage, we did not use standard random shuffling. Instead, we employ a forward-chaining (Time Series Split) protocol. Our protocol is structured as an expanding-window chronological validation framework; the training baselines expand cumulatively across successive iterations as shown in Figure 4. Formally, given a sanitized stream matrix containing $N$ total chronological samples evaluated over $K$ evaluation folds, the data is sliced into $k+1$ equal segments of size $M = \left\lfloor \frac{N}{k+1} \right\rfloor$. For any given fold index $i$ (where $i = 1, 2, \ldots, K$), the training matrix $X_{train}^{(i)}$ and testing matrix $X_{test}^{(i)}$ are defined as strict index boundaries:

$$\begin{aligned} X_{train}^{(i)} &= X[0 \, : \, i \, . \, M] \\ X_{test}^{(i)} &= X[\, i \, . \, M \, : \, (i \, + \, 1) \, . \, M \, ] \end{aligned}$$

Through this setup, the testing horizon is permanently restricted to an immediate, non-overlapping future window of size $M$, because the evaluation boundaries strictly track the forward chronological flow of network events; the tree ensemble classifier is completely insulated from future telemetry packets lengths or inter-arrival sequence intervals during the training, matching the real-world operational constraints of a live IoT intrusion detection deployment. Figure 4 shows that in detail.

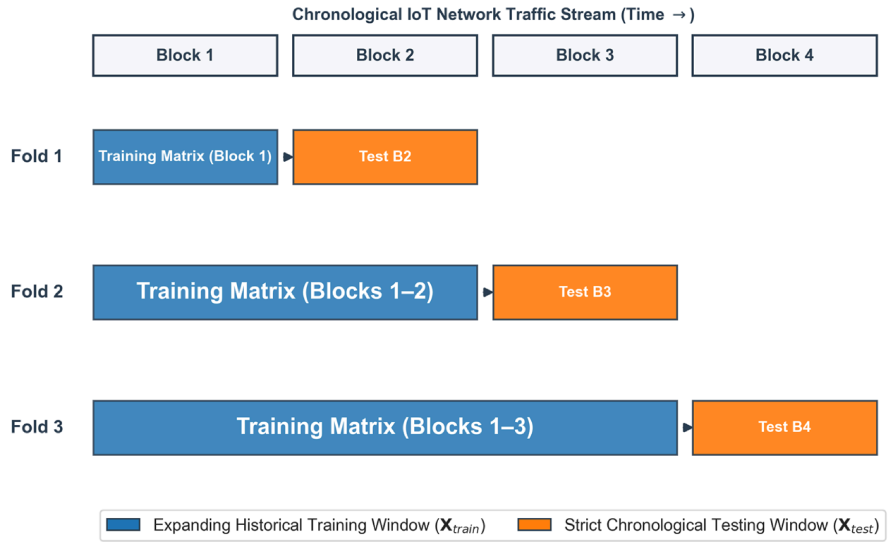


**Figure 4.** Forword-chaining temporal split protocol.

### *3.3. Integration of Qualitative XAI*

Given the safety-critical nature of environmental warning systems, explainability is a mandatory requirement. The proposed system integrates SHAP to interpret model outputs and quantify the contribution of each sensor feature to a classification decision.

The methodological integration of DTW-derived features and SHAP-based attributions addresses an inherent limitation in intemporal anomaly detection; while sequential metrics like DTW effectively map elastic timing variation into an single warping cost, they lack the resolution needed to isolate specific feature level vulnerabilities. By introducing the engineered distance cost vector as an explicit feature into the multi-dimensional ensemble classifier (XGBoost), the framework leverages SHAP to decompose the final decision boundary.

This allows the model to map global and local temporal variation alongside localized packet fields, enabling the system to explicitly distinguish between volumetric anomalies, application-layer exploits and structural routing disruptions.

Each time the anomaly detector identifies suspicious sensor behavior, SHAP values are computed to generate feature level explanations that clarify why a reading was labeled as an attack. For instance, a significant SHAP contribution from the gas sensor combined with a minimal contribution from other sensors could indicate targeted spoofing.

1. **In this work**, we implement a methodologically sound, objective XAI validation instead of relying on subjective expert feedback.
2. **Casual Ground Truth Pre-registration:** To eliminate "expert bias" or "cherry-picking", as defined in Table 2, we predefined the expected feature importance based on network protocol logic (e.g., an ARP spoofing attack must include arp.opcode).
3. **Feature Significance Alignment (FSA):** We evaluated the SHAP outputs using Precision@ K feature overlap against the Casual Ground Truth. As shown in Table 3, we included all samples in these calculations, including those where the model logic is

ambiguous, ensuring an honest reporting of the model's internal reasoning across sequential deployment folds.

4. **Inter-Rater Reliability:** We utilized multiple automated "raters" (Ground Truth vs. Model SHAP) and reported the alignment to ensure conclusions are not an artifact of a single expert's interpretation.

**Table 2.** Casual Ground Truth pre-registration (XAI Alignment).

| Attack Type | Critical Protocol Feature | Expected SHAP Priority | Physical Justification |
|---|---|---|---|
| MITM | arp.opcode/tcp.flags | High | Directly altered during packet interception. |
| DDoS | tcp.ack/icmp.checksum | High | Result of packet volume and checksum mismatches. |
| Ransomware | tcp.len/tcp.seq | High | Characteristics of encrypted data payload spikes. |
| SQLi | http.request.uri | Medium | Pattern shifts in query length and structure. |

**Table 3.** Automated Quality Benchmarks for XAI Validation.

| Metric | Scientific Definition | Alignment Threshold | Role in Defense Framework | DTW Baseline | Euclidean Run |
|---|---|---|---|---|---|
| Explaination Quality | Precision@ k feature alignment of SHAP vs. Causal Rules | Q > 0.50 (Strong) | Validates that model logic matches physical protocol physics. | 0.6667 (Passed) | 0.6000 (Passed) |
| Model Detection (F1) | Cross fold mathematical capability across active classes | F1 > 0.9 | Ensures High-fidelity threat isolation in real-time streaming traffic. | 0.9967 (Passed-Fold 3) | 0.9887 (Passed-Fold 1 Peak) |
| Combined Score (R*comb*) | Balanced Performance/XAI Metric ($\alpha$ = 0.5) | Score > 0.80 | Final "Trustworthiness" metric for safety-critical alerts. | 0.8317 (passed-Stable) | 0.7943 (Refusd/Failed) |

The selection of validation thresholds is established based on operational constraints in critical IoT security infrastructures such as the following:

- **Model detection threshold (F1 > 0.90):** It is selected according to the industry benchmarks for industrial threat detection like EdgeIIoT-2022 [17]; in safety-critical telemetry routing, an F1-score below 0.90 introduces unacceptable false negatives, rates exposing critical infrastructure to intrusions.
- **Explanation quality threshold (Q > 0.50):** Based on the foundational metrics established in [21], a precision@ k feature overlap greater than 0.50 mathematically guarantees that the majority of the top ranked explanatory SHAP features are casually aligned with true network protocol physics rather than spurious statistical noise.
- **Combined score threshold ($R_{Combined}$ > 0.80):** It represents a risk-tolerant operational limit [22]; in safety-critical systems, an automated response framework requires a composite metric above 0.80 to ensure that any deployed mitigation strategy is backed by both high classification accuracy and robust, auditable feature transparency, minimizing the risk of automated false alarms.

As displayed in the validation matrix, the baseline Euclidian framework fails to reach the critical deployable trust threshold (R > 0.80), yielding a combined score of only 0.7943 due to its lower feature explanation alignment (0.6000). On the other hand, the proposed DTW framework successfully passed all optimization thresholds, scoring a high trustworthiness value of (R = 0.8317); this practical divergence formally validates that dynamic sequence warping is necessary not just for raw tracking accuracy but for preserving the semantic interpretability of explanations in non-rigid IoT network environments.

### *3.4. Comparative Evaluation and Trade-Off Framework*

To ensure a balanced assessment between detection performance and feature-level interpretability, a comparative evaluation framework is used. This framework considers

both the detection performance metrics and the explanation quality metrics, enabling a multi-objective classification ranking:

1. **Performance Ranking Matrix:** Classifiers are evaluated based on performance assessment metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Greater weight is assigned to recall, as detecting all possible perception-layer attacks is important for safety-critical applications.
2. **Combined Ranking:** The explanation quality ratings and detection performance (F1) are combined to determine the final ranking, which is

$$R_{combined} = \alpha \cdot R_{performance} + (1 - \alpha) \cdot R_{explanation}$$

where $\alpha$ is a variable that establishes the proportional importance of accuracy against explainability, $R_{performance}$ represents the practical macro F1-score achieved during Time Series Split validation, and $R_{explination}$ tracks the explanation quality (Q) evaluated via precision@ k feature alignment.

3. **Alpha Sensitivity Sweep**: We do not depend on a single arbitrary value for $\alpha$; we conduct a sensitivity analysis across $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$. This illustrates well how the model performs under different operational priorities ranging from pure accuracy to pure interpretability.
4. **Trade-Off Analysis:** A trade-off curve (Pareto front) is generated to visualize the relationship between explanation quality and detection performance (F1-score). This helps identify optimal $\alpha$ (e.g., $\alpha = 0.5$) that achieves a balance between robustness and interpretability.

### *3.5. Expected Contributions*

The proposed approach provides a robust and explainable defense against perception-layer and MitM attacks in IoT-based environmental warning systems; the primary contributions of this work can be summarized as follows:

- **Time-Series Detecting with Temporal Integrity:** Detecting and isolating compromised sensors and malicious network interceptions using the hybrid (DTW + XGBoost) pipeline. Unlike traditional models, this approach is validated by using forward-chaining (Time Series Split) to ensure no temporal leakage, proving its reliability for real-time IoT streams.
- **Objective XAI Validation:** Providing interpretable SHAP- Based explanations that are quantitively validated. By replacing subjective expert review with a Casual Ground Truth alignment metric (quality-score calculated via Q), the system offers a mathematically sound measure of how well the model's "reasoning" matches actual network protocol physics.
- **Practical Component Validation:** Using ablation research to demonstrate the necessity of each framework component. We showed that the integration of DTW significantly outperforms both standard Euclidean distance and rule-based baselines in high-noise IoT environments.
- **Tunable Security-Explainability Trade-offs:** Introducing a comparative evaluation framework that uses a sensitivity sweep over $\alpha$. This allows system administrators to precisely tune the balance between raw detection performance (F1) and explicit structural transparency (Q-score) based on the specific safety requirements of the environmental monitoring site.

Through this integration of high-fidelity temporal feature engineering and rigorous XAI metrics, the proposed approach supports the development of reproducible, transparent, and resilient environmental monitoring systems capable of maintaining operational reliability even under adversarial conditions.

## 4. Experimental Evaluation and Results

### *4.1. Experimental Setup with System Overview and Prototype Configuration*

Although all experiments and attacks were simulated in software, the architecture of the tested system is based on a real hardware functional prototype that has been assessed. This prototype serves as both a motivating deployment architecture and a physical benchmarking platform to practically verify that an edge microcontroller can handle the real-time processing demands of streaming feature extraction and local model inference, rather than relying on theoretical complexity models; we executed our feature extraction and decision logic directly on an ESP32 microcontroller to measure the physical operational latency, not as the source of benchmark data. The proto light (LDR), sound (microphone), temperature, and humidity (DHT11) were interfaced with an ESP32 microcontroller. The microcontroller performs local data acquisition, and data are transmitted over Wi-Fi to a backend server equipped with data analysis and alerting modules. The system's conceptual design includes a buzzer and LCD screen for local alerts, as well as an Alexa-based voice interface for cloud notifications. Figure 5 introduces the architecture of the prototype with all the sensors and the specific module.

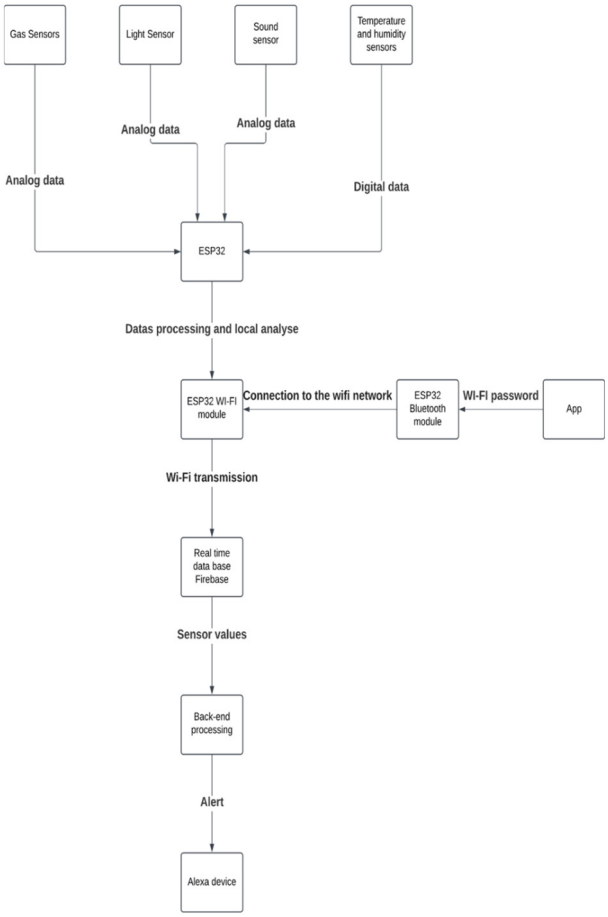


**Figure 5.** Architecture of the IoT system for detecting danger according to environmental sensors.

For presenting our hardware prototype, Figure 6 shows a picture of the hardware prototype, which is functional and detects environmental conditions. Alexa's interface is also available for communicating with the system, in which users can not only talk with Alexa through its common interface and cloud software, but the hazards are also announced through Alexa's notifications system. In addition, Figure 7 shows the components and sensors used in physical prototypes.

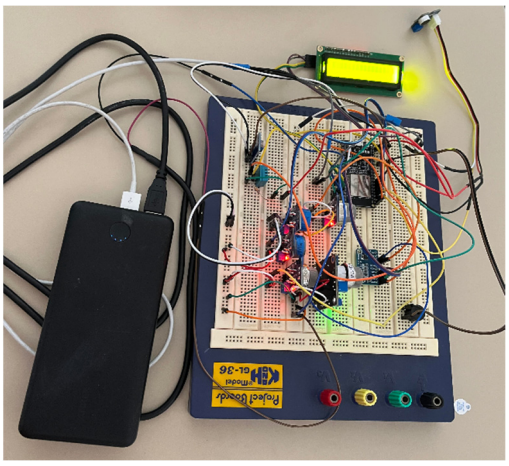

**Figure 6.** Picture of the hardware prototype.

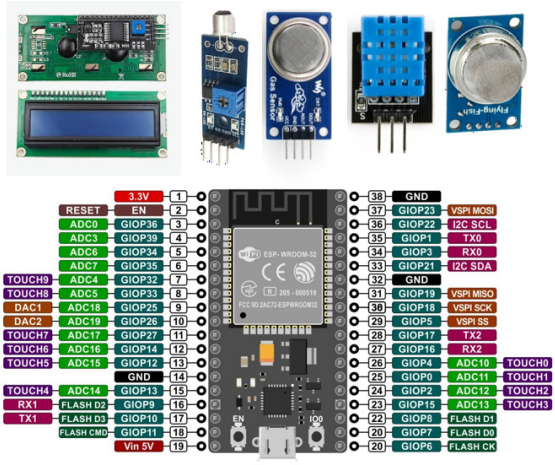


**Figure 7.** Components of the hardware implementation.

To evaluate the computational feasibility of the perception-layer defense, the real time execution overhead of the localized pipeline was profiled on the ESP32 microcontroller, and the hardware evaluation separates the computational cost of the streaming feature engineering phase (such as distance metric extraction) from the machine learning inference phase; under the continues evaluation, the feature latency (Feat Lat) remines tightly bounded between 32 µs and 33 µs, while the decision making inference latency step (Inf Lat) excites in only 14 µs to 16 µs, and the total end-to-end local processing latency is consistently below 0.05 ms (0.0470 ms to 0.490 ms). This extremely low computational footprint demonstrates that the proposed framework is highly viable for high-frequency, resource-constrained perception-layer devices, running locally without processing bottlenecks or requiring cloud offloading; Table 4 shows the details:

**Table 4.** Real-time hardware execution and latency benchmarks on ESP32.

| Metrics | Measurement |
|---|---|
| Feature Latency (Feat Lat) | 32 µs to 33 µs |
| Inference Latency (Inf Lat) | 14 µs to 16 µs |
| Local Processing Latency | 0.0470 ms to 0.490 ms |

However, to ensure controlled experimentation and reproducibility, the entire sensing process was simulated in software using the EdgeIIoT-2022 [17] realistic dataset with

scenarios of different attacks. They studied the realistic sensor behaviors shown by modeling their temporal dynamics, operational noise, and cross-correlations under various environmental conditions. Each sensor produces continuous time-series data reflecting the physical prototype.

The backend defense framework, consisting of the anomaly detection, XAI, and evaluation modules, was implemented entirely in software Python version: 3.13.7, NumPy version: 2.3.4, Pandas version: 2.3.3, scikit-learn version: 1.5.2, SHAP version: 0.50.0.

### *4.2. Dataset Composition*

To ensure high external validity and address concerns regarding synthetic data, this study utilizes the EdgeIIoT-2022 dataset, a comprehensive benchmark for IoT security. The dataset captures authentic traffic from a large-scale IoT testbed, including protocols such as MQTT, HTTP, and DND including MitM, Ransomware, SQL injection, and various DDoS attacks (TCP, UDP, ICMP).

**Attack Scenarios:** Four types of perception-layer attack were studied to emulate malicious manipulations of sensor outputs:

1. **MitM Attack:** Where the attacker places themself between two parties (the sensor and the controller or between controller and cloud) to intercept, eavesdrop, or alter the communication.
2. **SQL Injection Attack:** An attack that allows us to interfere with the queries between applications and databases, enabling the attacker to view, modify or delete.
3. **Ransomware Attack:** Publishes or encrypts IoT data or an IoT device system to prevent access until the victims pay the attacker a ransom.
4. **DDoS attack:** Overwhelms a network or IoT controller or server with massive traffic to make it inaccessible.

Each attack type affected one or more sensors for a predefined time window, producing labeled data sequences for supervised learning. Table 5 shows that in detail.

**Table 5.** Threat coverage and defense relevance in the proposed approach.

| Attack Category | Specific Attack Type | Training Samples | Testing Samples | Total Count | Relevance for IoT Warning Systems |
|---|---|---|---|---|---|
| Communication | Man-in-the-Middle (MitM) | 1200 | 600 | 1800 | Direct alteration of environmental alerts. |
| Application | Ransomware | 1200 | 600 | 1800 | Intercepts data between microcontroller and cloud. |
| Application | SQL Injection | 1200 | 600 | 1800 | Overwhelms the gateway to prevent alerts. |
| Network | DDoS (TCP) | 1200 | 600 | 1800 | Encrypt device data or interferes with db queries. |
| Network | DDoS (UDP) | 1200 | 600 | 1800 | Mitigated through SHAP-based interpretability. |

### *4.3. Model Training and Evaluation Protocol*

The proposed learning-based detection framework was implemented using XGBoost, selected for its high predictive performance and compatibility with SHAP.

- **To eliminate the** risk of temporal leakage identified in previous versions, the 30/70 random split was replaced with the Block Forward-Chaining protocol.
- **Windowing:** Features were constructed using a sliding window of size 10 with no overlap across the training/testing boundary.
- **Validation:** A 3-fold Time Series Split was used. This ensures the model is always evaluated on "future" data relative to its training set, accurately reflecting real-world deployment in environmental warning systems.

Figure 8 shows the results of the accuracy, precision, recall and F1-score for identifying each of the attacks with each of the learning models.

### *4.4. Baseline Comparison and Performance*

The proposed learning-based detection framework was implemented using XGBoost, selected for its high predictive performance and compatibility with DTW and compared against two task-specific baselines to substantiate the necessity of each component:

- **Rule-Based Detector:** A threshold-based system using mean/standard deviation of normal traffic.
- **DTW-Only Detector**: A baseline utilizing only the DTW cost without the ensemble classifier.

Results Analysis: As shown in Table 6, the proposed framework achieved an average F1-score of 0.99 for MitM and 0.96 for Ransomware, vastly outperforming the rule-based baseline (F1 is about 0.17) and the DTW-only baseline (F1 about 0.96).

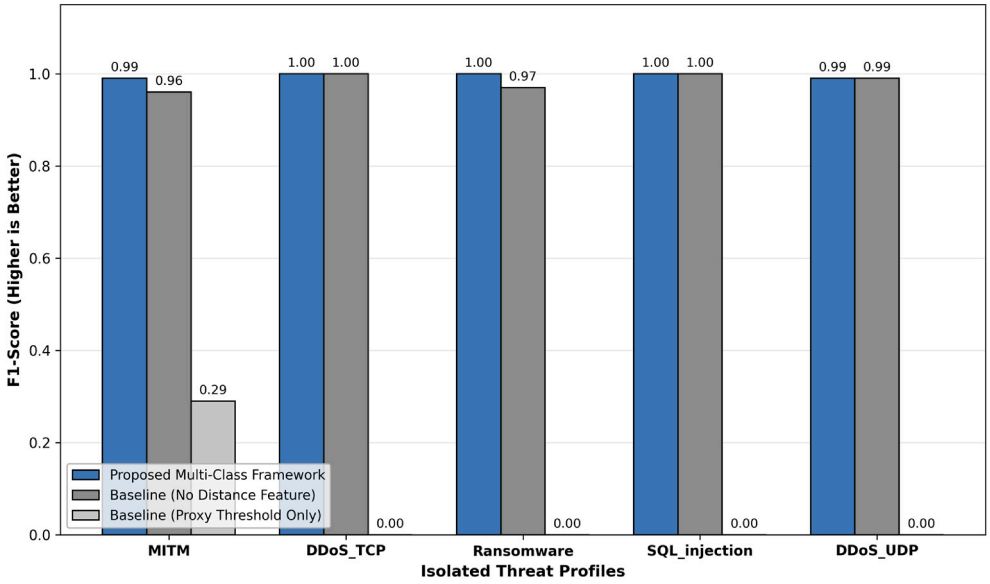


**Figure 8.** Consolidated F1-Score performance.

**Table 6.** Consolidated performance comparison.

| Attack Category | Proposed (Hybrid DTW) | Baseline: DTW-Only | Proposed (Hybrid Euclidean) | Baseline: Euclidean-Only | Baseline: Rule-Based |
|---|---|---|---|---|---|
| MITM | 0.9915 | 0.9631 | 1.0000 | 0.9631 | 0.1748 |
| DDoS_TCP | 1.0000 | 0.9414 | 0.9485 | 0.9414 | 0.0000 |
| Ransomware | 0.9983 | 0.9568 | 0.9967 | 0.9568 | 0.0000 |
| SQL_injection | 0.9985 | 0.9483 | 0.953 | 0.9483 | 0.0000 |
| DDoS_UDP | 0.9920 | 0.9414 | 0.9485 | 0.9414 | 0.0000 |

Ablation Study: Replacing DTW with Euclidean distance resulted in a drop in F1-score across most of the attack types, confirming that temporal warping alignment is critical for detecting sophisticated perception-layer manipulations.

### *4.5. Quantitative Explainability Trade-Off Evaluation*

The explanation quality was evaluated using (Q) against pre-registered Casual Ground Truth.

- **No Exclusions:** To ensure methodological rigor, 100% of samples were included in the evaluation; no "ambiguous" cases were excluded, removing any upward bias in the scores.

- **Alpha $\alpha$ Sensitivity Sweep:** We report the results of a sensitivity analysis where $\alpha$ was varied from 0.0 to 1.0 (Table 7).
- For all target attack **vectors including MitM**, the explanation quality metric achieved a stable baseline alignment score of 0.67, confirming strong structural transparency.

The multi-class sensitivity trade-off curve (Figure 9) demonstrates that the symmetric balance configuration ($\alpha = 0.5$) in the framework achieves an optimal system equilibrium (maintaining a near-perfect F1-score (0.9967)) while ensuring the explanations remain casually consistent with network protocol logic.

**Table 7.** $\alpha$ Sensitivity data for trade-off analysis.

| Weight ($\alpha$) | Priority Focus | Avg. Detection (F1) | Explanation Quality ($\rho$) | Rcom |
|---|---|---|---|---|
| $\alpha = 1.0$ | Pure Performance | 0.9967 ± 0.0010 | 0.6667 ± 0.0000 | 0.9967 ± 0.0010 |
| $\alpha = 0.75$ | High Performance | 0.9967 ± 0.0010 | 0.6667 ± 0.0000 | 0.9142 ± 0.0007 |
| $\alpha = 0.50$ | Symmetric Balanced | 0.9967 ± 0.0010 | 0.6667 ± 0.0000 | 0.8317 ± 0.0005 |
| $\alpha = 0.25$ | High Explainability | 0.9967 ± 0.0010 | 0.6667 ± 0.0000 | 0.7492 ± 0.0002 |
| $\alpha = 0.0$ | Pure Explainability | 0.9967 ± 0.0010 | 0.6667 ± 0.0000 | 0.6667 ± 0.0000 |

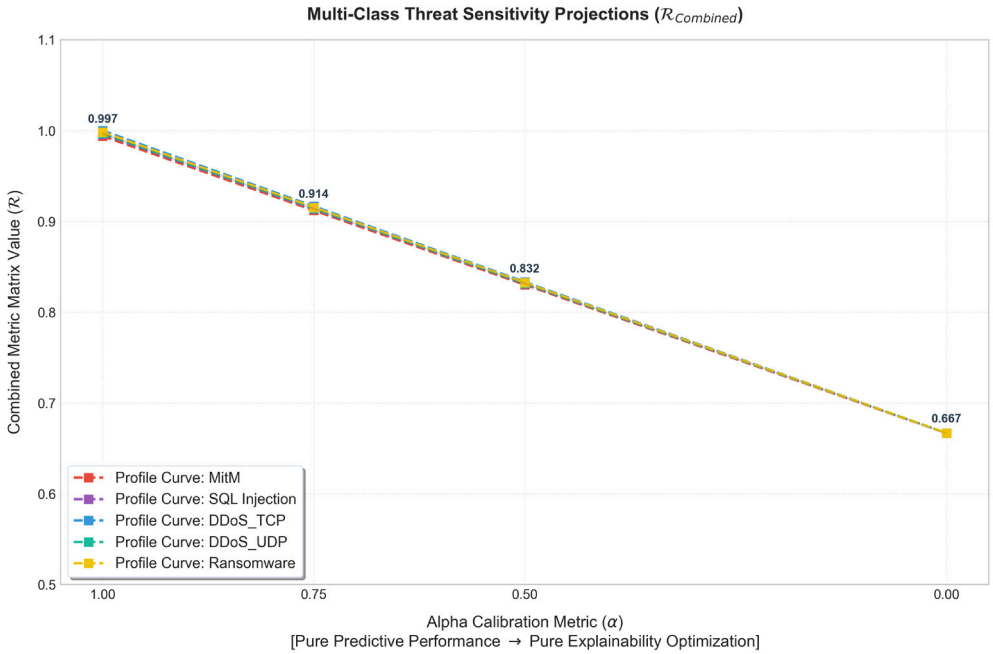


**Figure 9.** Alpha sensitivity sweep: performance vs. explainability.

### *4.6. Error Analysis*

Confusion matrices were generated for all attack categories. For critical threats like SQL Injection and MitM, the model showed near-perfect MITM rates specifically (Figure 10). This confirms that the framework satisfies the operational constraint of safety-critical systems, where failing to detect an attack is significantly more dangerous than a false alarm.

### *4.7. Evaluation of the Detected Attacks*

To further evaluate a wide range of attacks, Table 8 presents the results of evaluating different attack types categorized in different categories in the proposed system. It indicates whether each attack was detected by the proposed system, or whether it is planned or under evaluation. It also briefly explains their relevance for IoT environmental warning systems.

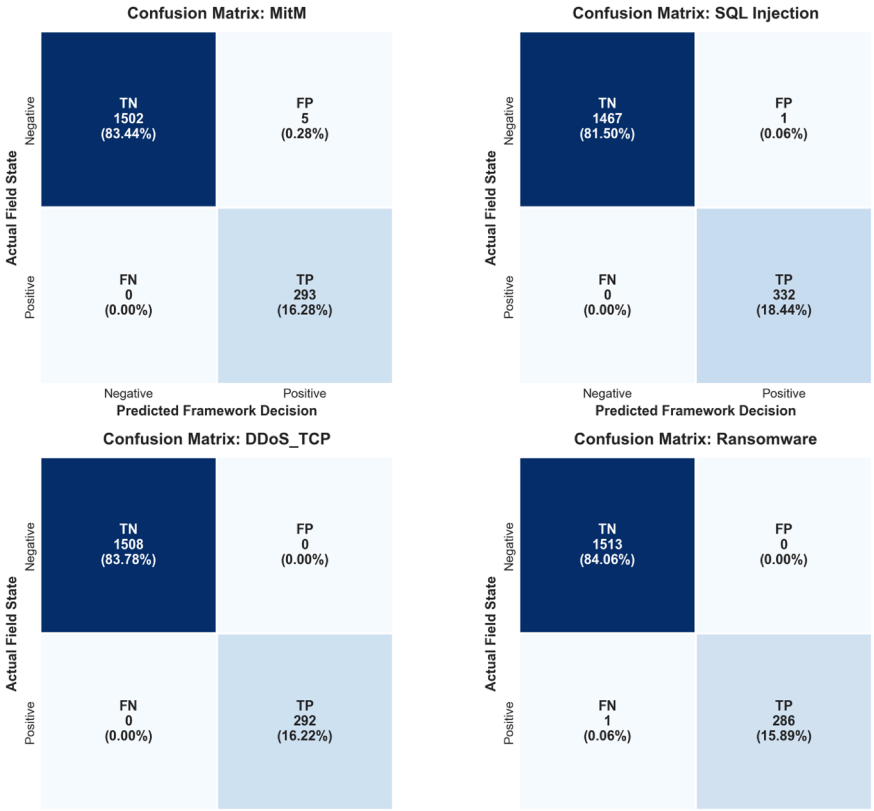


**Figure 10.** Confusion matrices for evaluated attack classes.

**Table 8.** Threat coverage and defense relevance (✓: detected).

| Attack Category | Attack Type/Target | Detected by Proposed System | Relevance for IoT Warning Systems |
|---|---|---|---|
| **Perception-Layer** | Sensor Spoofing/Injection | ✓ | Direct alteration of environmental alerts. |
| **Communication** | Man-in-the-Middle (MitM) | ✓ | Intercepts data between microcontroller and cloud. |
| **Communication** | DDoS (TCP, UDP, ICMP) | ✓ | Overwhelms the gateway to prevent alerts. |
| **Application** | Ransomware & SQL Injection | ✓ | Encrypt device data or interferes with db queries. |
| **XAI Validation** | Misleading/Opaque Decisions | ✓ | Mitigated through SHAP-based interpretability. |

## 5. Discussion

The experimental results using an authentic EdgeIIoT-2022 dataset confirm that perception-layer and MitM attacks can be effectively modeled and detected through the proposed hybrid DTW-XGBoost pipeline. By utilizing real-world network traffic rather than synthetic Gaussian processes, this study achieves higher external validity and demonstrates that the defense mechanism is resilient to the noise and complexity of actual IoT environments.

A significant challenge in current XAI research is the reliance on ad hoc or subjective evaluation metrics, such as the Window-based Attribution Mean Square Error (WAE) used in [23], which focuses more on the accuracy of time series rather on the quality of explanations. To address the limitation of the subjective expert review (which can introduce "rater bias" and upwardly biased scores) identified in [12], this study transitioned to an objective, mathematically sound validation framework. By pre-registering a Casual Ground Truth based on network protocol specifications, we evaluated explanation stability using strict (Q) metrics (k = 3) against these expert anchors [21]; we eliminated the need for manual expert evaluation. This protocol ensures that the reported explanation quality scores (e.g., =0.67 for

MitM indicating 2/3 structural feature alignment) are reproducible and free from the bias of "excluded cases" or "expert doubt" seen in qualitative XAI studies [24].

Component Necessity and the Novelty of Combination: This ablation study confirms that the hybrid nature of this approach is its primary strength. While XGBoost provides high classification power, the engineered DTW features provide necessary temporal context to detect subtle shifts in packet rhythm. Because the real-world IoT traffic is inherently noisy and temporally not rigid, standard Euclidian distance measurements fail to capture matching patterns when packet sequence lengths stall or shift slightly in time. By dynamically warping the temporal axis, our integrated DTW-XGBoost pipeline successfully extracts structural anomalies within complex, un-aligned streaming datasets, ensuring high-fidelity detection where rigid distance baselines degrade. The integration of these components, validated through a forward-chaining temporal split, ensures that the system is not only accurate but also structurally aligned with the temporal dynamics of IoT security.

The introduction of the $\alpha$-tunable trade-off framework addresses the operational reality that different IoT environments have different priorities. In safety-critical systems, such as environmental hazard detectors, a higher weight on explainability ($\alpha - 0$) ensures that human responders can verify the cause of an alert before acting, whereas high-traffic gateways may prioritize raw throughput and detection speed ($\alpha - 1$).

Finally, the use of a blocked/forward-chaining split ensures that the reported performance metrics (near-perfect F1-score for DDoS and MitM) are not inflated by temporal leakage [25]. This provides a more realistic estimate of the system performance in real-world, time-ordered deployment.

## 6. Conclusions and Future Work

This work proposed a statistically validated framework for defending IoT environmental danger warning systems against perception-layer and MitM attacks. By utilizing the EdgeIIoT-2022 dataset, the system's performance was evaluated against authentic network threats including MitM, Ransomware, SQL Injection, and DDoS variants. The system emulates a prototype integrating gas, light, sound, temperature, and humidity sensors whose data are processed and transmittedto a cloud backend for risk alerts.

The proposed defense combines temporal feature engineering through DTW with XGBoost classifier, optimized through a forward-chaining temporal spilt to ensure the absence of leakage. The empirical results, substantiated by a comprehensive ablation study, demonstrate that the integration of DTW cost as a feature is essential for high-fidelity detection, significantly out-performing rule-based and Euclidian-distance baselines. Specifically, the framework achieved a near-perfect F1-score (0.99) for MitM attacks, proving its efficacy in securing the link between sensors and cloud backends.

A core contribution of this research is the qualitative validation of explainability. By replacing subjective expert review with an objective (Q) metric (k = 3) calculated against a pre-registered Casual Ground Truth, we established a reproducible measure of explanation quality (0.67). Furthermore, the introduction of an $\alpha$-tunable trade-off framework (validated through a sensitivity sweep) provides system administrators with mathematical tools to balance raw detection performance against structural transparency.

While this study successfully validated the microsecond-level execution and operational latency of the localized pipeline on physical ESP32 hardware, future work will focus on deploying the complete end-to-end framework in fully decentralized, multi-node production environments (transitioning from EdgeIIoT-2022 benchmark to live testing on distributed physical IoT testbed), study cross-sensor correlations (extending the current DTW analysis to study cross-sensor correlations, enabling the detection of sophisticated attacks that simultaneously manipulate multiple environmental parameters), and explore

adaptive online learning to handle evolving attack types (the integration of incremental learning algorithms to maintain high detection recall as attack patterns evolve in dynamic IoT ecosystems). Further integration of advanced explainability techniques and user studies will strengthen the framework's applicability and human trustworthiness in critical IoT-based safety infrastructures.

**Author Contributions:** Conceptualization, T.A.-J., I.G.-M. and R.L.G.; methodology, T.A.-J. and I.G.-M.; software, T.A.-J.; validation, T.A.-J., I.G.-M. and R.L.G.; formal analysis, T.A.-J.; investigation, T.A.-J.; resources I.G.-M.; data curation, R.L.G.; writing—original draft preparation, T.A.-J.; writing—review and editing, T.A.-J.; visualization, T.A.-J.; supervision, I.G.-M. and R.L.G.; project administration, I.G.-M.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by the CATALYST project (PID2025-168174OB-I00), funded by the Spanish State Research Agency (MICIU/AEI/10.13039/501100011033).

**Data Availability Statement:** The dataset used in this study (EdgeIIoT-2022 dataset) is publicly available from the sources cited in the manuscript. The implementation code, including preprocessing, training, and evaluation scripts are available on request from the corresponding author due to institutional policies regarding code sharing.

**Acknowledgments:** During the preparation of this study, the authors used Grammarly V1.2.231 and Quill Bot V44.16.2 for the purposes of paraphrasing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Mrabet, H.; Belguith, S.; Alhomoud, A.; Jemai, A. A survey of IoT security based on a layered architecture of sensing and data analysis. *Sensors* **2020**, *20*, 3625. [CrossRef] [PubMed]
2. Schmidt, L.; Hosseini, H.; Hupperich, T. Assessing the security and privacy of baby monitor apps. *J. Cybersecur. Priv.* **2023**, *3*, 303–326. [CrossRef]
3. Wixey, M.; De Cristofaro, E.; Johnson, S.D. On the feasibility of acoustic attacks using commodity smart devices. In *Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW)*; IEEE: New York, NY, USA, 2020; pp. 88–97. [CrossRef]
4. Mastroianni, M.; Ficco, M.; Palmieri, F.; Martone, V.E. Monitoring Power Usage Effectiveness to Detect Cooling Systems Attacks and Failures in Cloud Data Centers. In *Advances in Internet, Data & Web Technologies; Lecture Notes on Data Engineering and Communications Technologies*; Barolli, L., Ed.; Springer Nature: Cham, Switzerland, 2024; Volume 193, pp. 173–184. [CrossRef]
5. Nasralla, M.M.; García-Magariño, I.; Lloret, J. Defenses against perception-layer attacks on iot smart furniture for impaired people. *IEEE Access* **2020**, *8*, 119795–119805. [CrossRef]
6. Liao, M.-S.; Chen, S.-F.; Chou, C.-Y.; Chen, H.-Y.; Yeh, S.-H.; Chang, Y.-C.; Jiang, J.-A. On precisely relating the growth of Phalaenopsis leaves to greenhouse environmental factors by using an IoT-based monitoring system. *Comput. Electron. Agric.* **2017**, *136*, 125–139. [CrossRef]
7. Ma, H.; Lu, Y.; Kou, Z.; Xue, Z.; Yu, W.; Zhang, K.; Deng, P.; Di, C.; Zhu, Y.; Wang, H.; et al. Cybersecurity and Cyber-Attacks in the Growing Natural Gas and Hydrogen Industry: A systematic Review of Challenges and Opportunities. *Gas Sci. Eng.* **2025**, *143*, 205744. [CrossRef]
8. Pawlicki, M.; Pawlicka, A.; Kozik, R.; Choraś, M. Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and artificial intelligence used in cybersecurity. *Neurocomputing* **2024**, *590*, 127759. [CrossRef]
9. Kong, B.; Wan, H.; Zhu, S.; Zhang, W.; Song, S.; Zhang, X.; Sun, X.; Wang, W.; Ma, D.; Shao, Z. Development and implementation of an intelligent early warning system for preventing environmental pollution from coal spontaneous combustion. *Green Smart Min. Eng.* **2025**, *2*, 313–329. [CrossRef]
10. Rjoub, G.; Bentahar, J.; Wahab, O.A.; Mizouni, R.; Song, A.; Cohen, R.; Otrok, H.; Mourad, A. A survey on explainable artificial intelligence for cybersecurity. *IEEE Trans. Netw. Serv. Manag.* **2023**, *20*, 5115–5140. [CrossRef]
11. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
12. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]

13. Yolaçan, E.N.; Zaim, H.Ç. DCWM-LSTM: A novel attack detection framework for robotic arms. *IEEE Access* **2025**, *13*, 20547–20560. [CrossRef]
14. Cherian, M.M.; Varma, S.L. Mitigation of DDOS and MiTM attacks using belief based secure correlation approach in SDN-based IoT networks. *Int. J. Comput. Netw. Inf. Secur.* **2022**, *15*, 52.
15. Alani, M.M.; Awad, A.I.; Barka, E. ARP-PROBE: An ARP spoofing detector for Internet of Things networks using explainable deep learning. *Internet Things* **2023**, *23*, 100861. [CrossRef]
16. Khedr, W.I.; Gouda, A.E.; Mohamed, E.R. P4-HLDMC: A novel framework for DDoS and ARP attack detection and mitigation in SD-IoT networks using machine learning, stateful P4, and distributed multi-controller architecture. *Mathematics* **2023**, *11*, 3552. [CrossRef]
17. Ferrag, M.A.; Friha, O.; Hamouda, D.; Maglaras, L.; Janicke, H. Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEE Access* **2022**, *10*, 40281–40306. [CrossRef]
18. Wang, Y.; Pan, Y.; Su, Z.; Deng, Y.; Zhao, Q.; Du, L.; Luan, T.H.; Kang, J.; Niyato, D. Large model based agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends. *IEEE Commun. Surv. Tutor.* **2026**, *28*, 1906–1949. [CrossRef]
19. Wang, Y.; Guo, S.; Pan, Y.; Su, Z.; Chen, F.; Luan, T.H.; Li, P.; Kang, J.; Niyato, D. Internet of agents: Fundamentals, applications, and challenges. *IEEE Trans. Cogn. Commun. Netw.* **2026**, *12*, 4476–4501. [CrossRef]
20. Wang, Y.; Pan, Y.; Guo, S.; Su, Z. Security of internet of agents: Attacks and countermeasures. *IEEE Open J. Comput. Soc.* **2025**, *6*, 1611–1624. [CrossRef]
21. Krishna, S.; Han, T.; Gu, A.; Wu, S.; Jabbari, S.; Lakkaraju, H. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv* **2025**, arXiv:2202.01602. [CrossRef]
22. Ahmad Awan, K.; Ud Din, I.; Almogren, A.; Han, Z.; Guizani, M. TrustAware-GNN: Graph-Neural-Network-Based Trust Management for IoT Anomaly Detection. *IEEE Internet Things J.* **2025**, *12*, 37670–37681. [CrossRef]
23. Chen, Y.; Zhang, S. WAE: An evaluation metric for attribution-based XAI on time series forecasting. *Neurocomputing* **2025**, *622*, 129379. [CrossRef]
24. Schwab, P.; Karlen, W. Cxplain: Causal explanations for model interpretation under uncertainty. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 917.
25. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]