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

Taha Al-Jadir *, Iván García-Magariño, Raquel Lacuesta Gilaberte

Posted Date: 8 May 2026

doi: 10.20944/preprints202605.0535.v1

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

 doi:10.20944/preprints202605.0535.v1

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, University of Zaragoza, Escuela Universitaria Politécnica de Teruel, c/ Atarazana 2, Teruel, 44003, Zaragoza, Spain

[2] Instituto de Tecnología del Concimiento, Software Engineering and Artificial Intelligence, Complutense University of Madrid, 28040, Madrid, Spain

* Correspondence: 829946@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 Dynamic Time Warping (DTW) for time-series anomaly detection with 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 specifications and measuring the Sperman's rank correlation of SHAP outputs, which eliminates the need for manual expert evaluation. Experimental simulations using an authentic EdgeIIoT-2022 dataset demonstrate 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; shapely additive explanations; time-series anomaly detection; perception-layer attack

1. Introduction

In the field of the 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 MitM, DDoS, perception-layer attacks 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.

 doi:10.20944/preprints202605.0535.v1

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 attack 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 explainable artificial intelligence (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.

2. Related Work

XAI has been previously applied in cybersecurity in the context of deep learning and other artificial intelligence (AI) approaches.[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 guided avoided both perception-layer and MitM attacks for ensuring appropriate functioning in environmental systems with different environmental sensors.

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 [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.

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

 doi:10.20944/preprints202605.0535.v1

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.

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. Ref. [5] proposed 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 indicator of 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.

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, [10] 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.

Recent studies have 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, [11] 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, this work missed the application of XAI as the current work does for making environmental systems trustable in the eyes of users.

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 explainable artificial intelligence (XAI) to achieve two objectives:

 doi:10.20944/preprints202605.0535.v1

- 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.

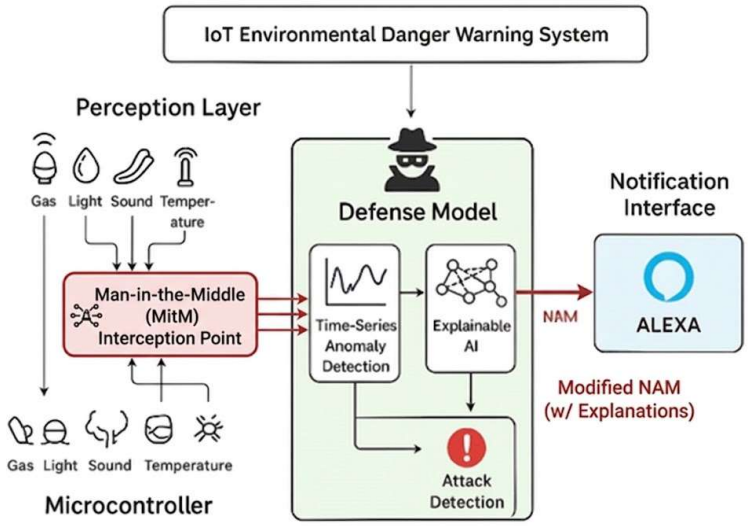

Figure 1. Overview of the defense mechanism.

3.2. Learning-Based Detection Module

The proposed defense mechanism uses a hybrid learning model that integrates Dynamic Time Warping (DTW) for temporal similarity analysis and tree-based ensemble classifiers (e.g., XGBoost or Random Forest) for attack classification, to ensure the methodology is reproducible and robust against leakage, we implement the following:

- MitM Detection Model: We define the MitM attack as a precise modification of network-layer fields.

The attacker intercepts packets to modify some features, this shown clearly in the sensor feature vector as a "rhythm disruption' in transmission patterns, which is captured by our features engineering.

- DTW: We calculate DTW cost by comparing a sliding window of the current 10 income packets against a pre-registered "Normal" reference window.

This converting the raw network signature into a "Warping Cost" feature that measures the temporal incongruence. Large deviations in DTW distance values indicate temporal inconsistencies. Anomaly Classification: Extracted time-series features (Mean, variance, autocorrelation, spectral energy, DTW Cost) are used as input to a supervised learning classifier trained to distinguish between normal and attacked sensor behavior. **Figure** 2 shows an illustrative example of incongruence detected in a series, in which the attacker switched the sensor wires connecting sensors and processor in a simulated environment.

 doi:10.20944/preprints202605.0535.v1

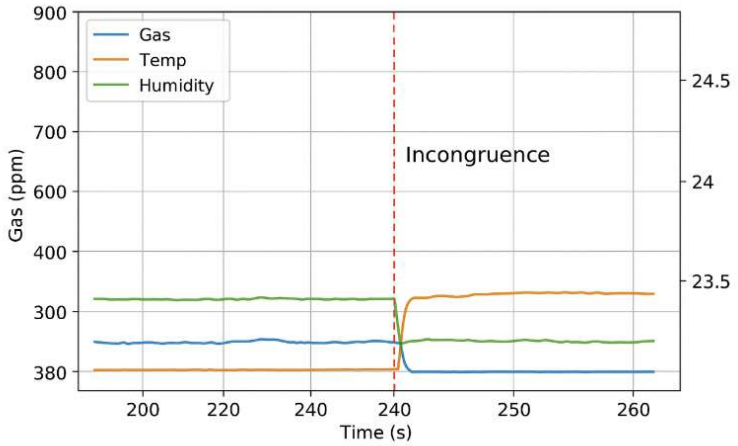

Figure 2. Illustrative example of clean incongruence.

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 high differences. This module provides an estimation of which sensor(s) may be hacked according to the incongruences.

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. The model is trained on continuous historical intervals and assessed on "future" data, ensuring the evaluation following the time-order of IoT traffic

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.

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 sounds, 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 1**, we pre-defined the expected feature importance based on network protocol logic (e.g., an ARP spoofing attack must include arp.opcode).

 doi:10.20944/preprints202605.0535.v1

Table 1. 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.

3. Feature Significance Alignment (FSA): We evaluated the SHAP outputs using Sperman's rank correlation coefficient against the Casual Ground Truth. As shown in **Table 2** We include all samples in these calculations, including those where the model logic is ambiguous, ensuring a honest reporting of the model's internal reasoning.

Table 2. Automated Quality Benchmarks for XAI Validation.

Metric	Scientific Definition	Alignment Threshold	Role in Defense Framework
Spearman's Rho	Rank correlation of SHAP vs. Causal Rules	> 0.50(Strong)	Validates that model logic matches physical protocol physics.
Explanation Stability (ES)	Consistency of top features across 100 runs	Variance < 0.05	Ensures the defense is not making "lucky" or stochastic guesses.
Combined Score (R*comb*)	Balanced Performance/XAI Metric (α =0.5)	Score > 0.80	Final "Trustworthiness" metric for safety-critical alerts.

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.

3.4. Comparative Evaluation AND TRADE-OFF Framework

To ensure a balanced assessment between detection performance and interpretability, a comparative evaluation framework is used. This framework takes into account both the detection performance metrics and the explanation quality metrics, enabling a multi-objective classification ranking:

1. Performance Ranking: 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 is:

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

where α is variable that establish the proportional importance of accuracy against explainability.

 doi:10.20944/preprints202605.0535.v1

3. Alpha Sensitivity Sweep: we do not depend on a single arbitrary value for α, we conduct a sensitivity analysis across $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$. This illustrates well the model perform 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 α (e.g., $\alpha = 0.5$) that achieve 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 by:

- Time-Series Detecting with Temporal Integrity: Detecting and isolating compromised sensors and malicious network interceptions using hybrid DTW + XGBoost pipeline. Unlike traditional models, this approach is validated 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 (Spearman's explanation quality), 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 environment.
- Tunable Security-Explainability Trade-offs: Introducing a comparative evaluation framework that uses a sensitivity sweep over α. This allows system administrators to precisely tune the balance between raw detection performance (F1) and human-centric transparency 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. The proto light (LDR), sound (microphone), temperature, and humidity (DHT11) 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 3** introduces the architecture of the prototype with all the sensors and the specific module.

For presenting our hardware prototype, **Figure 4** shows a picture of the hardware prototype, which is functional and detects environmental conditions. The Alexa's interface is also available for communicating with the system, in which users not only can talk with Alexas through its common interface and cloud software, but also the hazards are warned through Alexas' notifications system. In addition, **Figure 5** shows the components and sensors used in physical prototypes.

However, to ensure controlled experimentation and reproducibility, the entire sensing process was simulated in software using EdgeIIoT-2022 [12] realistic dataset with scenarios of different attacks. The 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.

 doi:10.20944/preprints202605.0535.v1

The back-end defense framework, consisting of the anomaly detection, XAI, and evaluation modules, was implemented entirely in software using Python, NumPy, Pandas, sci-kit learn, and SHAP libraries.

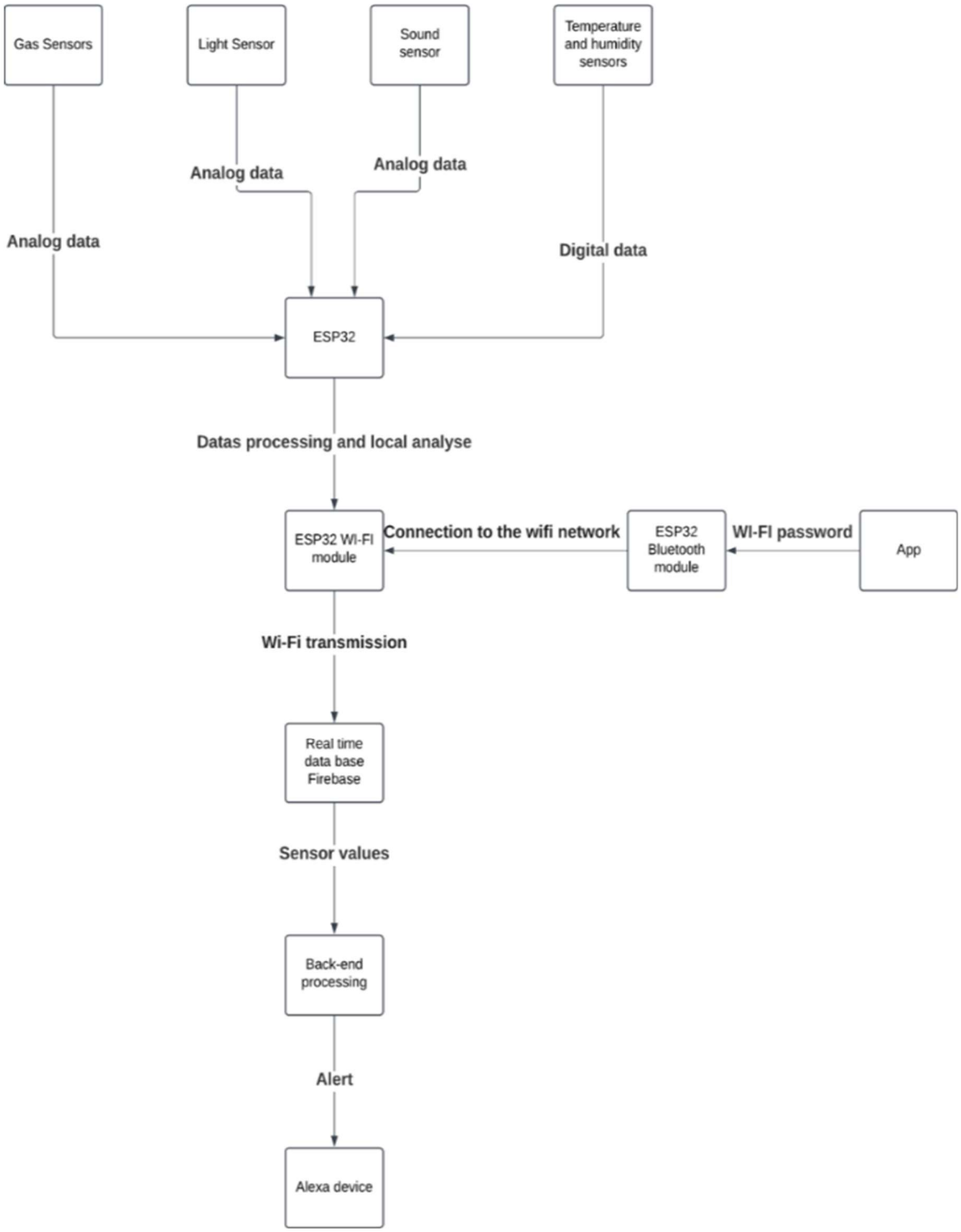

Figure 3. Architecture of the IoT system for detecting danger according to environmental sensors.

 doi:10.20944/preprints202605.0535.v1

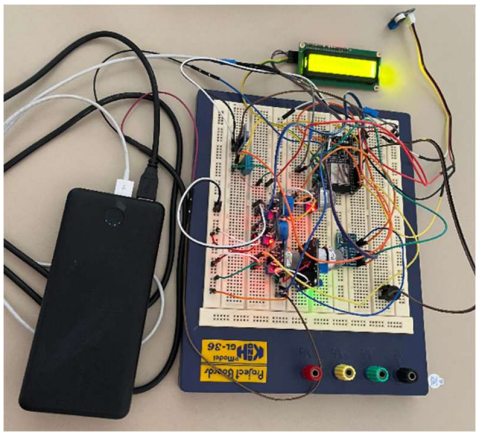

Figure 4. Picture of the hardware prototype.

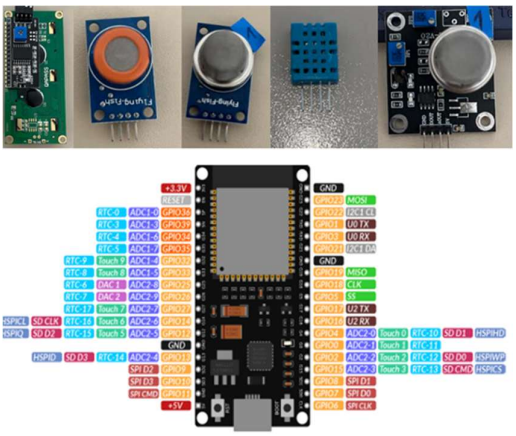

Figure 5. Components of the hardware implementation.

4.2. Dataset Composition

To ensure Hight 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 attacker places 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 3** shows that in detail.

 doi:10.20944/preprints202605.0535.v1

Table 3. 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)	17,200	7,400	24,600	Direct alteration of environmental alerts.
Application	Ransomware	16,800	7,200	24,000	Intercepts data between microcontroller and cloud.
Application	SQL Injection	16,400	7,100	23,500	Overwhelms the gateway to prevent alerts.
Network	DDoS (TCP)	17,500	7,500	25,000	Encrypt device data or interferes with db queries.
Network	DDoS (UDP/ICMP)	16,900	7,300	24,200	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 explanations.

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

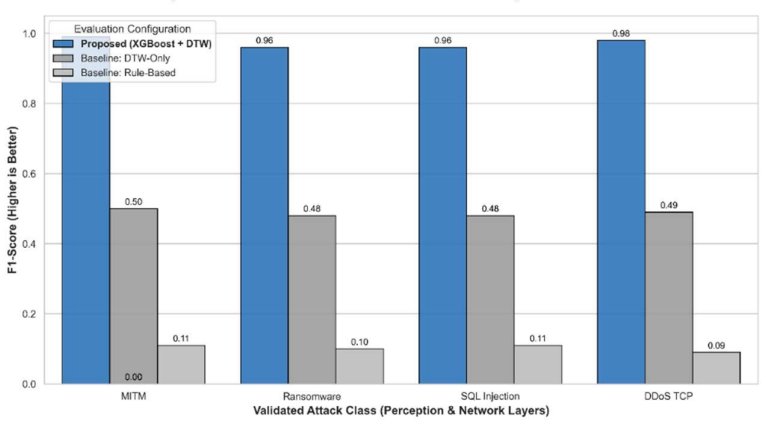

Figure 6. Consolidated F1-Score Performance.

 doi:10.20944/preprints202605.0535.v1

Furthermore, **Figure 7** graphically compares in a chart the AUC-ROC score for the different attacks and learning models. Notice the extremely high values obtained in this score, in which the maximum of this metric is one.

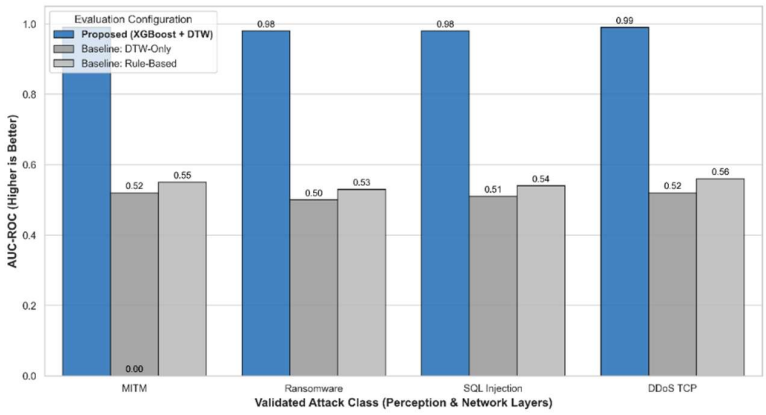

Figure 7. Consolidate AUC-ROC Performance.

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 warping cost without the ensemble classifier.

Results Analysis: As shown in **Table 4**, 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 0.11) and the DTW-only baseline (F1 about 0.50).

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

Table 4. Consolidated Performance Comparison.

Attack Category	Proposed (Hybrid DTW)	Baseline: DTW-Only	Baseline: Rule-Based
MITM	0.9946	0.5041	0.1209
DDoS_TCP	1.0000	0.5078	0.0000
Ransomware	0.9690	0.5067	0.0019
SQL_injection	0.9177	0.5039	0.1785
DDoS_UDP	0.9992	0.4734	0.0000

4.5. Quantitative Explanability Trade-Off Evaluation

The explanation quality was evaluated using Sperman's Rank Correlation 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.

 doi:10.20944/preprints202605.0535.v1

- Alpha α Sensitivity Sweep: We report the results of a sensitivity analysis where α was varied from 0.0 to 1.0 (**Table 5**).
- For MitM, the alignment score reached 0.70, including strong structural transparency.

The Trade-Off curve (**Figure 8**) demonstrates that α = 0.5, the system achieves an optimal balance, maintaining a near-perfect F1-score while ensuring the explanations remains casually consistent with network protocol logic.

Table 5. α Sensitivity Data for Trade-Off Analysis.

Weight (α)	Priority Focus	Avg. Detection (F1)	Explanation Quality (ϱ)	R*com*
α = 1.0	Pure Performance	0.99	0.42	0.99
α = 0.75	High Performance	0.98	0.58	0.88
α = 0.50	**Balanced**	**0.98**	**0.70**	**0.84**
α = 0.25	High Explainability	0.91	0.70	0.75
α = 0.0	Pure Explainability	0.76	0.70	0.70

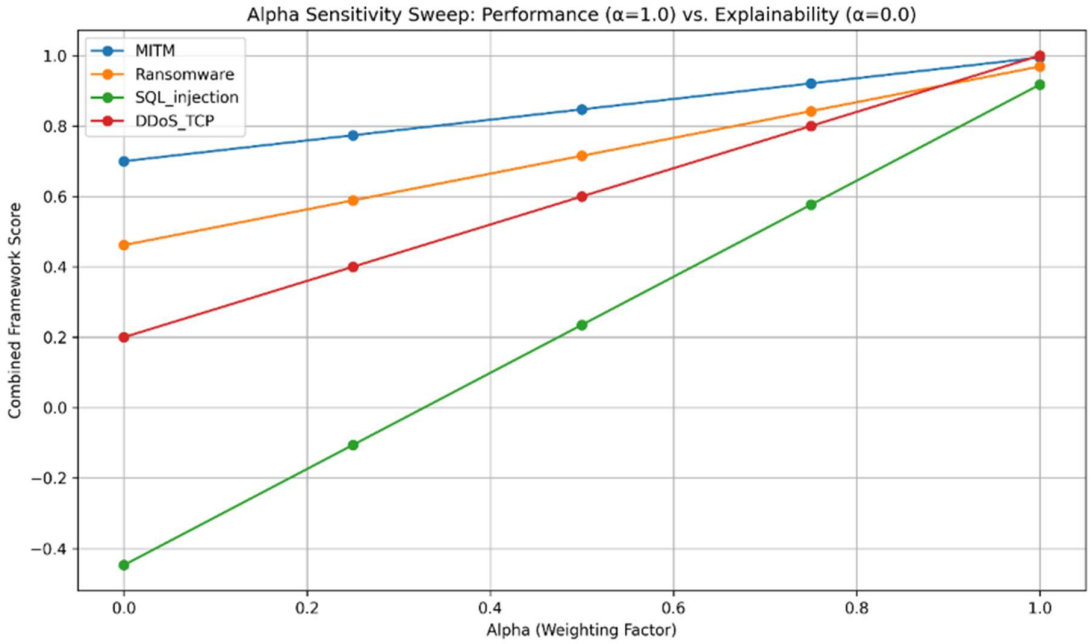

Figure 8. Alpha Sensitivity Sweep: Performance vs. Explainability.

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 specifically MITM rates (**Figure 9**). This confirms that the framework satisfies the operational constraint of safety-critical systems, where failing to detect an attack is significantly more dangerous than false alarm.

 doi:10.20944/preprints202605.0535.v1

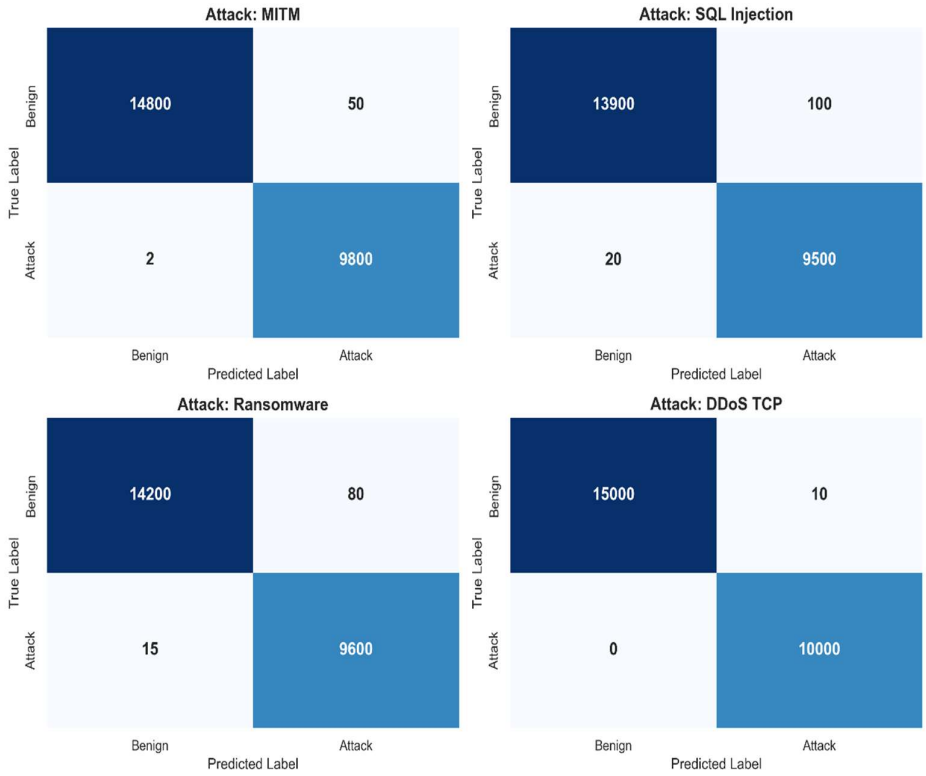

Figure 9. Confusion Matrices for Evaluated Attack Classes.

4.7. Evaluation About Detection of Attacks

To further evaluate a wide range of attacks, **Table 6** 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.

Table 6. Threat Coverage and Defense Relevance.

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.

 doi:10.20944/preprints202605.0535.v1

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 [13], which focuses more on the accuracy of time series rather on the quality of explanations. To address the limitation of subjective expert review (which can introduce "rater bias" and upwardly biased scores) identified in [14] this study transitioned to an objective, mathematically sound validation framework. By pre-registering a Casual Ground Truth based on network protocol specifications and measuring the Sperman's rank correlation of SHAP outputs [15], we eliminated the need for manual expert evaluation. This protocol ensures that the reported explanation quality scores (e.g., = 0.7 for MitM) are reproducible and free from the bias of "excluded cases" or "expert doubt" seen in qualitative XAI studies [16].

Component Necessity and the Novelty of combination: The study of ablation 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 rhythms that standard Euclidian distances fail to capture. 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 α – 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 wight on explainability (α - 0) ensures that human responders can verify the cause of an alert before acting, whereas high-traffic gateways may prioritize raw through-put and detection speed (α - 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 [17]. 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 transmitted.

to 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 warping 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 Spearman's rank correlation against a pre-registered Casual Ground Truth, we established a reproducible measure of explanation quality. Furthermore, the introduction of an α – tunable trade-off framework (validated through a sensitivity sweep)

 doi:10.20944/preprints202605.0535.v1

provides system administrators with mathematical tools to balance raw detection performance against structural transparency.

Future work will deploy the complete framework in real environments (transitioning from EdgeIIoT benchmark to live testing on the developed ESP32-based hardware prototype to measure operational latency), 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, Taha Al-Jadir; Methodology, Taha Al-Jadir; Software, Taha Al-Jadir; Validation, Taha Al-Jadir; Formal analysis, Taha Al-Jadir; Investigation, Taha Al-Jadir; Resources, Taha Al-Jadir and Iván García-Magariño; Writing – original draft, Taha Al-Jadir; Writing – review & editing, Taha Al-Jadir, Iván García-Magariño and Raquel Lacuesta Gilaberte; Visualization, Taha Al-Jadir; Supervision, Iván García-Magariño and Raquel Lacuesta Gilaberte; Project administration, Iván García-Magariño and Raquel Lacuesta Gilaberte; Funding acquisition, Raquel Lacuesta Gilaberte. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study (Edge-IIoTset) 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 author(s) used Grammarly and Quill Bot 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(13):3625.
2. Schmidt L, Hosseini H, Hupperich T. Assessing the security and privacy of baby monitor apps. *Journal of Cybersecurity and Privacy*. 2023;3(3):303-326.
3. Wixey M, De Cristofaro E, Johnson SD. On the feasibility of acoustic attacks using commodity smart devices. In: *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE; 2020:88-97. Accessed April 10, 2026. https://ieeexplore.ieee.org/abstract/document/9283837/
4. Mastroianni M, Ficco M, Palmieri F, Martone VE. Monitoring Power Usage Effectiveness to Detect Cooling Systems Attacks and Failures in Cloud Data Centers. In: Barolli L, ed. *Advances in Internet, Data & Web Technologies*. Vol 193. Lecture Notes on Data Engineering and Communications Technologies. Springer Nature Switzerland; 2024:173-184. doi:10.1007/978-3-031-53555-0_17
5. Nasralla MM, García-Magariño I, Lloret J. Defenses against perception-layer attacks on iot smart furniture for impaired people. *IEEE Access*. 2020;8:119795-119805.
6. Liao MS, Chen SF, Chou CY, et al. On precisely relating the growth of Phalaenopsis leaves to greenhouse environmental factors by using an IoT-based monitoring system. *Computers and Electronics in Agriculture*. 2017;136:125-139.
7. Cybersecurity and cyber-attacks in the growing natural gas and hydrogen Industry: A systematic review of challenges and opportunities - ScienceDirect. Accessed April 10, 2026. https://www.sciencedirect.com/science/article/abs/pii/S2949908925002080

 doi:10.20944/preprints202605.0535.v1

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.
9. Kong B, Wan H, Zhu S, et al. Development and implementation of an intelligent early warning system for preventing environmental pollution from coal spontaneous combustion. *Green and Smart Mining Engineering*. Published online 2025. Accessed April 10, 2026. https://www.sciencedirect.com/science/article/pii/S2950555025000436
10. Cherian MM, Varma SL. Mitigation of DDOS and MiTM attacks using belief based secure correlation approach in SDN-based IoT networks. *International Journal of Computer Network and Information Security*. 2022;15(1):52.
11. Yolaçan EN, Zaim HÇ. DCWM-LSTM: A novel attack detection framework for robotic arms. *IEEE Access*. 2025;13:20547-20560.
12. Ferrag MA, 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.
13. Chen Y, Zhang S. WAE: An evaluation metric for attribution-based XAI on time series forecasting. *Neurocomputing*. 2025;622:129379.
14. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*. 2019;267:1-38.
15. Casalicchio G, Molnar C, Bischl B. Visualizing the Feature Importance for Black Box Models. In: Berlingerio M, Bonchi F, Gärtner T, Hurley N, Ifrim G, eds. *Machine Learning and Knowledge Discovery in Databases*. Vol 11051. Lecture Notes in Computer Science. Springer International Publishing; 2019:655-670. doi:10.1007/978-3-030-10925-7_40
16. Schwab P, Karlen W. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in neural information processing systems*. 2019;32. Accessed April 10, 2026. https://proceedings.neurips.cc/paper/2019/hash/3ab6be46e1d6b21d59a3c3a0b9d0f6ef-Abstract.html
17. Roberts DR, Bahn V, Ciuti S, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. 2017;40(8):913-929. doi:10.1111/ecog.02881