

XAI-IDS: An Explainable Intrusion Detection System for IoT/IIoT Networks Using SHAP and LLM

Zohra Chair*, Issam Jabri*, Sonia Ben Brahim[†], and Hanen Lajnef[†]

**IResCoMath Laboratory, National Engineering School of Gabès (ENIG), University of Gabès, Gabès, Tunisia*

Emails: chair.zohra@gmail.com, issam.jabri@enig.rnu.tn

[†]InnoV'COM Laboratory, Higher School of Communications of Tunis (SUP'COM), University of Carthage, Ariana 2083, Tunisia

Emails: hanen.lajnef@enetcom.usf.tn, sonia.benbrahim@supcom.tn

Abstract—The growing deployment of IoT and IIoT systems has intensified the need for transparent intrusion detection. While machine learning-based IDS achieve high accuracy, their black-box nature limits adoption in Security Operations Centers (SOC). The proposed framework employs XGBoost for both attack detection and SHAP-based feature attribution, unifying the detection and explanation pipeline into a single model. LLaMA 3.1 8B Instant, deployed through the Groq API, transforms SHAP outputs into structured, human-readable security alerts. Experimental results, evaluated on the Edge-IIoTset dataset, demonstrate that the proposed framework produces highly faithful and actionable explanations, achieving a Faithfulness score of 0.76, a Feature Mention Rate of 0.94, an Actionability Score of 0.99, and a BERTScore F1 of 0.84. A preliminary evaluation using simulated SOC analyst profiles further suggests a positive perceived quality (overall score of 4.21/5.0), pending validation through studies with certified human analysts. The results demonstrate that combining XAI and LLMs significantly improves the transparency and interpretability of AI-driven IDS, representing a promising step toward operationally usable explainable intrusion detection for IoT and IIoT environments.

Index Terms—Intrusion Detection System, Explainable Artificial Intelligence, SHAP, Large Language Models, IoT Security, IIoT Security.

I. INTRODUCTION

The rapid proliferation of Internet of Things (IoT) and Industrial Internet of Things (IIoT) devices has significantly expanded the attack surface of modern digital infrastructures. Traditional Intrusion Detection Systems (IDSs), based on signature matching or predefined thresholds, struggle with the diversity and sophistication of contemporary attacks. Consequently, Machine Learning (ML) and Deep Learning (DL) have become increasingly important for intrusion detection because of their ability to learn complex attack patterns and generalize to previously unseen threats [1]–[7].

Despite their effectiveness, most ML- and DL-based IDSs remain black-box models, limiting insight into the reasoning behind their predictions. This lack of transparency can reduce analyst trust and complicate incident investigation in Security Operations Centers (SOCs). Explainable Artificial Intelligence (XAI), particularly SHAP (SHapley Additive exPlanations),

addresses this limitation by quantifying the contribution of individual features to model predictions [8]–[12]. However, SHAP outputs are primarily numerical and may require substantial cybersecurity expertise to translate them into actionable operational insights.

Large Language Models (LLMs) provide a potential solution by transforming technical security information into human-readable reports and recommendations. Recent studies have investigated their use for cyber threat intelligence, alert triage, and security reporting [13], [14]. However, their use as primary intrusion detectors remains limited by classification performance and inference costs [15]–[17]. Hybrid approaches therefore combine conventional ML models for detection with LLMs for contextualized explanations. Examples include HuntGPT, which integrates ML, SHAP, and conversational AI for threat hunting [18], and eX-NIDS, which uses cyber threat intelligence to augment LLM prompts for network-flow explanations [19].

Nevertheless, existing XAI-LLM approaches provide limited quantitative assessment of explanation faithfulness and stability, rarely measure the alignment between model-derived evidence and LLM-generated explanations, and largely overlook human-centered evaluation for SOC adoption [20], [21]. To address these gaps, we propose XAI-IDS, a framework integrating XGBoost-based intrusion detection, SHAP-based feature attribution, and LLM-driven alert generation for IoT/IIoT environments. The framework is evaluated on the Edge-IIoTset dataset [22]. The main contributions of this work are:

- A unified framework integrating intrusion detection, SHAP-based explainability, and LLM-assisted alert generation for IoT/IIoT networks.
- A methodology for transforming SHAP feature attributions into structured and actionable cybersecurity explanations using LLM prompting techniques.
- A quantitative evaluation framework combining explainability metrics (Sparsity, Consistency, and Faithfulness) with alert-quality metrics (FMR, Actionability Score, ROUGE-L, and BERTScore) and human-centered assessment.

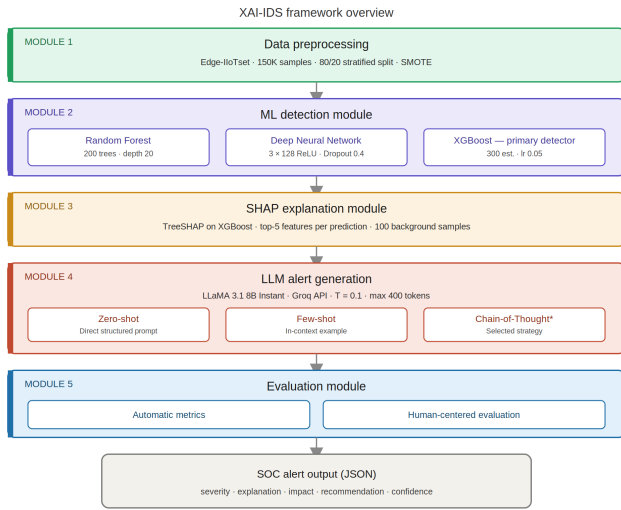


Fig. 1. Proposed IDS-XAI-LLM architecture.

- An experimental evaluation on Edge-IIoTset demonstrating the effectiveness of the proposed XAI-LLM pipeline.

II. PROPOSED SYSTEM ARCHITECTURE

The XAI-IDS pipeline consists of five interconnected modules, as illustrated in Fig. 1. Each module is designed to be independent and replaceable, ensuring modularity and reproducibility.

A. Data Preprocessing

The proposed framework is evaluated on the Edge-IIoTset dataset [22], a comprehensive benchmark for IoT and IIoT intrusion detection. To reduce computational cost while preserving class distribution and attack diversity, a stratified subset of 150,000 samples is extracted from the original dataset. Data preprocessing includes the removal of duplicate, missing, and infinite records, followed by the elimination of 18 non-informative features (e.g., IP addresses, timestamps, protocol-specific identifiers, and MQTT payload fields). Such feature reduction improves efficiency while maintaining detection performance [6], [11]. Categorical attributes are encoded using one-hot encoding, while attack labels are label-encoded. The dataset is then divided into training and testing sets using an 80/20 stratified split. Numerical features are standardized using StandardScaler fitted on the training data only, and class imbalance is addressed through SMOTE [23] with $k=3$, applied exclusively to the training set.

This preprocessing pipeline ensures data quality, balanced class representation, and reproducible experimental conditions for intrusion detection and explainability analysis.

B. ML Detection Module

The machine learning detection module is responsible for identifying and classifying network attacks. To provide a comprehensive evaluation, three complementary classifiers are trained and compared: Random Forest (RF), Deep Neural

Network (DNN), and XGBoost. These models were selected due to their widespread adoption in IoT and IIoT intrusion detection research and their demonstrated effectiveness on large-scale cyber security datasets [4]–[6].

The RF classifier is implemented as an ensemble of 200 decision trees with a maximum depth of 20 and balanced subsample weighting to mitigate class imbalance. Owing to its robustness, interpretability, and compatibility with TreeSHAP explanations, RF serves as a strong baseline for intrusion detection and explainability analysis [8].

The DNN classifier consists of three fully connected hidden layers with 128 ReLU neurons each. Batch Normalization, Dropout (0.4), and L2 regularization ($\lambda = 10^{-4}$) are applied after each layer to improve generalization. The network is optimized using Adam with a learning rate of 10^{-3} and trained for up to 15 epochs using a batch size of 256. Early Stopping (patience=4) and ReduceLROnPlateau scheduling (factor=0.5, patience=2) are employed to prevent overfitting and improve convergence. Deep learning architectures have demonstrated strong effectiveness in modeling complex and non-linear attack patterns in IoT and IIoT traffic [4], [5].

The XGBoost classifier is configured with 300 estimators, a maximum tree depth of 6, and a learning rate of 0.05. To enhance generalization, both row subsampling (subsample=0.8) and feature subsampling (colsample_bytree=0.8) are employed. Furthermore, L1 (reg_alpha=0.1) and L2 (reg_lambda=1.0) regularization are incorporated to reduce model complexity and prevent overfitting. For multiclass intrusion detection, the model uses the `multi:softprob` objective function and is optimized using the multiclass logarithmic loss (`mlogloss`) evaluation metric. Gradient boosting approaches such as XGBoost have consistently achieved state-of-the-art performance in network intrusion detection due to their ability to capture complex feature interactions while maintaining efficient inference [5], [7].

To ensure full reproducibility, a fixed random seed (`seed=42`) is applied throughout the experimental pipeline, including data sampling, preprocessing, and model training. Model performance is evaluated using Accuracy, Precision, Recall, and weighted F1-score.

C. SHAP Explanation Module

To enhance the transparency of intrusion detection decisions, the proposed framework incorporates SHAP (SHapley Additive exPlanations), a game-theoretic explainability technique that quantifies the contribution of each input feature to a model prediction [23], [24].

SHAP explanations are computed directly on the XGBoost detection model, ensuring full consistency between the reported detection decision and its corresponding explanation — an important methodological requirement absent from prior dual-model XAI-IDS designs, where a separate, potentially divergent model is used solely for explanation purposes. TreeExplainer is employed to efficiently compute exact Shap-

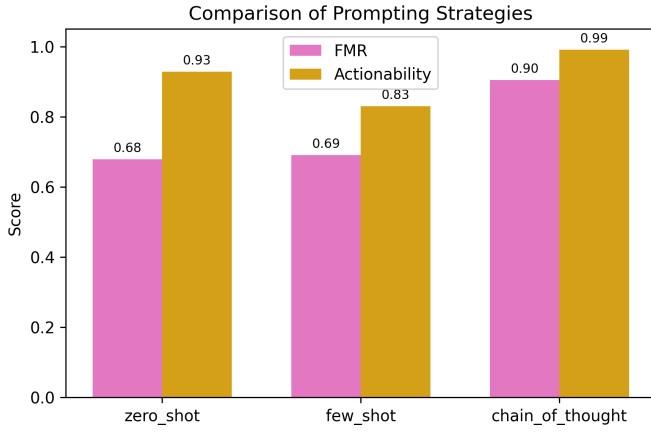


Fig. 2. Prompting Strategies comparison

key values using a representative background sample of 100 training instances.

For each prediction, the five features with the highest absolute SHAP values are extracted and ranked according to their contribution magnitude. The resulting explanation is represented as

$$F = \{(f_i, v_i, \phi_i, d_i)\}_{i=1}^5,$$

where f_i denotes the feature name, v_i the observed feature value, ϕ_i the absolute SHAP importance score, and $d_i \in \{+, -\}$ indicates whether the feature contributes positively or negatively to the predicted attack class.

The generated SHAP explanations constitute the primary input to the subsequent LLM-based alert generation module. In addition, explanation quality is assessed using three complementary metrics, namely Sparsity, Consistency, and Faithfulness [25].

D. LLM-Based Alert Generation Module

The LLM module transforms SHAP explanations into human-readable SOC alerts and mitigation recommendations. Three prompting strategies were evaluated: Zero-Shot, Few-Shot, and Chain-of-Thought (CoT). Preliminary experiments indicated that CoT achieved the highest SHAP-LLM alignment and generated the most actionable explanations (Fig. 2). Therefore, it was adopted for the final evaluation.

Each prompt contains the predicted attack type, the detection confidence score, source IP address, protocol information, and the five most influential SHAP features ranked by importance. For each feature, the prompt provides its name, observed value, SHAP contribution score, and contribution direction. To ensure explicit utilization of explainability information, the model is instructed to directly reference the SHAP-selected features when generating explanations and recommendations.

The generated output is constrained to a structured JSON format containing five fields: *severity*, *explanation*, *impact*, *recommendation*, and *confidence note*. LLaMA 3.1 8B Instant is deployed through the Groq API using a low-temperature

configuration ($T=0.1$) and a maximum output length of 400 tokens. Samples classified as *Normal* were excluded because they do not require analyst intervention or mitigation actions.

E. Evaluation Module

The proposed framework is evaluated at three levels: intrusion detection performance, explanation quality, and alert generation quality.

Detection performance is assessed using Accuracy, Precision, Recall, and F1-score.

The quality of SHAP explanations is evaluated using three complementary criteria [24]. Sparsity measures the degree to which explanatory importance is concentrated among a small number of features. Consistency measures the reproducibility of SHAP explanations by computing the Jaccard similarity of top-k features between two independent runs on identical test samples. To further assess robustness beyond this deterministic baseline, Consistency is additionally evaluated under small Gaussian input perturbations, measuring the Jaccard similarity between the top-5 features of the original and perturbed explanations. Faithfulness measures the correspondence between the identified important features and the model's actual decision process through feature-ablation analysis. These metrics jointly capture the interpretability, stability, and reliability of the generated explanations.

The quality of LLM-generated alerts is evaluated using Feature Mention Rate (FMR), Actionability Score (AS), ROUGE-L, and BERTScore F1. FMR measures the proportion of top SHAP features explicitly referenced in generated explanations. AS quantifies the presence of operational mitigation actions by detecting predefined cyber security action verbs. ROUGE-L and BERTScore F1 evaluate lexical and semantic similarity with reference alerts, respectively. In addition, average latency and token consumption are reported to assess computational efficiency.

Finally, a human-centered evaluation is performed using five simulated SOC analyst profiles, following the human-centered XAI evaluation taxonomy of Kim et al. [20] and the functionally-grounded evaluation framework of Doshi-Velez and Kim [21]. Generated alerts are rated according to Clarity, Accuracy, Actionability, Completeness, and Trust on a five-point Likert scale using severity-calibrated Gaussian priors.

III. EXPERIMENTAL EVALUATION

A. Experimental Setup

All experiments were conducted on the Edge-IIoTset dataset [22], a realistic IoT/IIoT cybersecurity benchmark collected from a real-world testbed comprising multiple IoT devices, including Raspberry Pi, Arduino, and ESP32 platforms. The dataset contains network traffic generated through several industrial and IoT protocols, including MQTT, Modbus, ARP, DNS, and HTTP, and covers 15 traffic categories consisting of one normal class and 14 attack classes, including DDoS variants, ARP Spoofing, SQL Injection, Cross-Site Scripting (XSS), DNS Spoofing, Backdoor, Password Cracking, Port Scanning, and Vulnerability Scanning.

To maintain computational efficiency while preserving class distributions, a stratified subset of 150,000 instances was extracted from the original dataset. This subset was used for model training and detection evaluation, ensuring robust performance estimation across all traffic classes. SHAP-based explainability analysis, by contrast, is conducted on a smaller stratified sample, as exhaustive per-instance SHAP computation across the full 150,000-sample set is computationally prohibitive and unnecessary given the convergence behavior demonstrated below. SHAP explanations were computed on a stratified sample of 285 test instances (up to 20 per class, capped by class availability), including 265 model-predicted attack instances across all 14 attack categories. This choice was empirically validated through a convergence analysis: Sparsity was recomputed across increasing per-class sample sizes (3 to 20), varying by only 0.0042 across the entire range and by 0.0013 between $N=15$ and $N=20$ per class, confirming that the metric had stabilized well within the chosen sample size. LLM alert quality metrics (FMR, Actionability, ROUGE-L, BERTScore) are evaluated over a further stratified subsample of 96 attack alerts (7 per class, 5 for MITM) selected from the 265 attack instances, balancing coverage across all 14 attack categories with API cost and latency.

All machine learning models were implemented in Python using Scikit-learn, TensorFlow, and XGBoost. The XGBoost classifier was selected as the primary intrusion detection model due to its superior classification performance. For explainability, SHAP values were computed directly on the XGBoost model using TreeExplainer, ensuring that the generated explanations are fully consistent with the detection decisions reported by the framework. SHAP explanations were generated using the SHAP library. Natural-language alert generation was performed using LLaMA 3.1 8B Instant accessed through the Groq API. To ensure reproducibility, all experiments were conducted with a fixed random seed of 42. The experiments were executed in the Kaggle cloud environment equipped with an NVIDIA Tesla T4 GPU with 16 GB of VRAM.

B. Experimental Results

1) *Intrusion Detection Performance*: Table I shows that XGBoost achieved the best overall performance, with 98.29% accuracy, 98.35% weighted F1-score, 98.47% precision, and 98.29% recall. RF achieved comparable results, whereas DNN performed lower across all metrics. These results support the selection of XGBoost as the primary detector.

TABLE I
COMPARATIVE PERFORMANCE OF THE EVALUATED MODELS

Model	Accuracy	F1-Score	Precision	Recall
Random Forest	0.9825	0.9831	0.9843	0.9825
DNN	0.9487	0.9447	0.9639	0.9487
XGBoost	0.9829	0.9835	0.9847	0.9829

2) *Evaluation of SHAP Explanations*: The Sparsity score was 0.6133, reflecting genuine heterogeneity in explanation

concentration across attack types: volumetric flooding attacks (DDoS_ICMP, DDoS_UDP) and Normal traffic exhibit highly concentrated feature signatures (Sparsity ≈ 0.82 – 0.83), whereas reconnaissance and application-layer attacks (Fingerprinting, Port Scanning, SQL Injection) rely on a broader combination of protocol and payload characteristics, yielding more distributed explanations (Sparsity ≈ 0.50 – 0.57).

Consistency reached a perfect 1.0000 under identical repeated runs confirming TreeSHAP’s deterministic behavior on XGBoost; however, under small Gaussian input perturbations ($\sigma=0.01$, 5 trials per sample, $N=30$ stratified subsample), Consistency dropped to 0.6187, revealing sensitivity of tree-based split thresholds to minor input changes despite this deterministic same sample baseline.

Feature-ablation experiments, conducted on 265 model-predicted attack instances using the top-5 SHAP features showed that removing these features reduced model confidence by a factor of $3.11\times$ compared to removing 5 randomly selected features, yielding a normalized Faithfulness score of 0.7564.

TABLE II
QUALITY ASSESSMENT OF GENERATED EXPLANATIONS

Metric	Value
Sparsity (normalized Shannon entropy)	0.6133
Consistency (Jaccard, same-sample runs)	1.0000
Consistency (Jaccard, Gaussian noise $\sigma=0.01$)*	0.6187
Faithfulness (SHAP vs random)	0.7564

* $N=30$ stratified subsample (2/class) due to computational cost

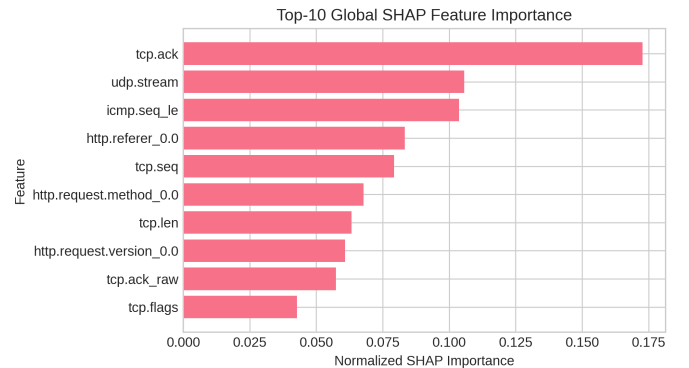


Fig. 3. Top-10 global feature importance obtained from normalized SHAP values. The figure highlights the network traffic attributes that contribute most to the intrusion detection decisions.

Figure 3 presents the global SHAP feature importance aggregated across 265 explained attack samples. The results show that only a limited number of features dominate the model decisions, which is consistent with the sparsity score reported in Table II. This observation indicates that the proposed IDS relies on a compact set of highly informative network attributes while maintaining high predictive performance.

3) *Evaluation of LLM-Generated Alerts*: Table III summarizes the LLM alert quality metrics. The proposed framework

achieved a Feature Mention Rate (FMR) of 0.941, indicating strong incorporation of SHAP-derived evidence into generated alerts across all attack types, and an Actionability Score of 0.997, showing that nearly all alerts contained actionable mitigation recommendations. ROUGE-L reached 0.095, reflecting moderate lexical overlap with reference alerts given the paraphrastic nature of LLM-generated cybersecurity text. BERTScore F1 reached 0.842, indicating strong semantic similarity between generated alerts and reference templates when both explanation and recommendation fields are considered. The average generation latency was 6585 ms, consistent with near real-time post-detection SOC triage workflows. Per-component analysis shows XGBoost (0.050 ± 0.004 ms/sample) and SHAP (7.2 ms/sample) contribute negligible overhead, with end-to-end latency dominated by LLM inference (6585 ± 1697 ms) over the Groq API. The high variance reflects network and API load rather than pipeline instability.

TABLE III
EVALUATION OF LLM-GENERATED EXPLANATIONS AND RECOMMENDATIONS

Metric	Value
Feature Mention Rate	0.9410
Actionability Score	0.9974
ROUGE-L	0.09449
BERTScore F1	0.8427

4) *Human-Centered Evaluation*: Table IV reports the results of the simulated human evaluation resulting in a total of 480 evaluations ($96 \text{ alerts} \times 5 \text{ experts}$). The generated explanations achieved an overall score of 4.21/5.0. Accuracy obtained the highest score (4.29/5.0), suggesting that the explanations closely reflect the underlying model predictions and SHAP-based evidence. Completeness (4.20/5.0), Clarity (4.21/5.0), and Actionability (4.21/5.0) also received strong ratings. Trust obtained the lowest score (4.14/5.0), remaining above the 4.0 threshold.

The low standard deviation (± 0.24) indicates high agreement among the simulated expert profiles, and results are consistent with the automatic evaluation metrics (FMR = 0.941, Actionability = 0.997), supporting the complementarity of automatic and human-centered evaluation.

TABLE IV
SIMULATED HUMAN EVALUATION OF GENERATED EXPLANATIONS

Criterion	Score (/5)
Clarity	4.21
Accuracy	4.29
Actionability	4.21
Completeness	4.20
Trust	4.14
Overall Score	4.21

5) *Discussion*: These results confirm the effectiveness of the proposed pipeline; the following situates them relative to existing literature.

Despite the strong detection performance achieved by existing approaches on the Edge-IIoTset dataset [4]–[7], the integration of explainable AI techniques and natural-language alert generation remains largely unexplored. Among related XAI-LLM frameworks, [18] combines SHAP/LIME with GPT-3.5 but relies on the outdated KDD99 dataset and evaluates explanation quality only through linguistic readability formulas, while eX-NIDS [19] enhances LLM explanations via CTI-based prompt augmentation rather than model-intrinsic explainability. Neither framework evaluates alignment between feature-attribution explanations and LLM-generated reports, and human-centered evaluation remains largely unexplored.

In contrast, the proposed framework introduces three complementary evaluation dimensions, namely Feature Mention Rate (FMR), Actionability Score, and Human Expert Assessment, providing a more comprehensive validation of explainable intrusion detection systems. This multi-layered design is deliberate: FMR and Actionability are lexical proxy metrics assessing feature coverage and the presence of predefined action oriented verbs, respectively without guaranteeing semantic correctness or technical relevance, so generic recommendations could in principle score highly on these alone. BERTScore and the human-centered evaluation (Section III-B-4) were included to jointly assess semantic quality and perceived operational usefulness. As shown in Table V, existing XAI-LLM studies mainly demonstrate feasibility but provide limited quantitative assessment of explanation quality and model alignment.

6) *Limitations*: Despite the promising results obtained, several limitations of the present study should be acknowledged.

First, the human evaluation relies on severity-calibrated proxy scoring rather than real SOC analyst judgment, which represents the primary validation gap of this work and is prioritized for future real-world application-grounded evaluation [19].

Second, LLM hallucination risk is reduced but not eliminated by SHAP-grounded inputs, FMR verification, and structured JSON constraints; RAG integration with verified knowledge bases (MITRE ATT&CK, CVE) is a promising mitigation strategy.

Third, the LLM alert generation module was applied only to samples classified as attacks. Consequently, the framework's behavior on benign traffic, false positives, and borderline cases—an important scenario in real SOC triage—remains untested and is identified as a direction for future work.

Fourth, Faithfulness approaches zero for two classes (DDoS UDP, Vulnerability scanner) due to near-saturated prediction confidence, a metric limitation, not unfaithful explanations while the remaining twelve classes score 0.68–0.88

IV. CONCLUSION AND FUTURE WORK

This paper presented XAI-IDS, an explainable framework that bridges the gap between high-performing black-box detection and operational decision support in IoT/IIoT networks. By unifying XGBoost classification, TreeSHAP feature attribution, and LLaMA 3.1 Chain-of-Thought prompting, our system

TABLE V
COMPARISON OF EXPLAINABILITY AND LLM EVALUATION METHODOLOGIES

Work	Faithfulness	Consistency	FMR	Actionability	Human Evaluation
HuntGPT [18]	×	×	×	×	Limited
eX-NIDS [19]	×	×	×	×	×
XAI-IDS (Ours)	✓	✓	✓	✓	✓

successfully transforms complex network logs into structured, actionable SOC alerts. Experimental evaluation on the Edge-IIoTset dataset demonstrated high detection accuracy, perfect consistency, and a Feature Mention Rate (FMR) of 0.94, proving that the LLM reliably grounds its reasoning in SHAP-derived empirical evidence with minimal hallucinations.

Future work will explore five key directions: (i) local quantized LLM deployment on edge devices to reduce latency and API dependency; (ii) integration with SIEM platforms for operational SOC deployment; (iii) evaluation of lightweight transformer-based intrusion detectors against XGBoost; (iv) comparison of alternative LLMs under different computational constraints; and (v) dataset expansion to include emerging attack categories and more diverse operational scenarios.

REFERENCES

- [1] A. J. G. de Azambuja, C. Plesker, K. Schützer, R. Anderl, B. Schleich, and V. R. Almeida, "Artificial Intelligence-Based Cyber Security in the Context of Industry 4.0—A Survey," *Electronics*, vol. 12, no. 8, p. 1920, 2023.
- [2] Y. Sanjalawe, S. Fraihat, S. Al-E'mari, and S. N. Makhadmeh, "A Review of Artificial Intelligence-Based Intrusion Detection in Industrial Internet of Things," *Discover Internet of Things*, vol. 6, no. 44, 2026.
- [3] N. Kalpani and N. Rodrigo, "Securing Industry 4.0: A Systematic Review of AI-Driven Intrusion Detection Approaches and Emerging Trends," *Journal of Reliable Intelligent Environments*, vol. 12, 2026.
- [4] S. Gupta, "AI-Driven Intrusion Detection in Internet of Things Networks Using the Edge-IIoTset Dataset," *International Journal of Emerging Research in Engineering and Technology*, vol. 7, no. 1, pp. 180–187, 2026, doi: 10.63282/3050-922X.IJERET-V7I1P124.
- [5] A. J. Frhan, "Hybrid Intrusion Detection with Edge-IIoTset Dataset," *International Journal of Education and Learning Systems*, vol. 11, pp. 1–12, 2026.
- [6] T. Hasan, A. Hossain, M. Q. Ansari, and T. H. Syed, "Enhanced Intrusion Detection in IIoT Networks: A Lightweight Approach with Autoencoder-Based Feature Learning," in *Proc. 10th Int. Conf. Internet of Things, Big Data and Security (IoTBDs)*, Porto, Portugal, 2025, pp. 620–627, doi: 10.5220/0013203700003944.
- [7] W. Gao, M. Wang, Y. Pei, F. Li, and C. Wang, "A Lightweight Multi-Classification Intrusion Detection Model for Edge IoT Networks," *MDPI Electronics*, 2025.
- [8] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.
- [9] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating Standard Feature Sets Towards Increased Generalisability and Explainability of ML-based Network Intrusion Detection," *arXiv preprint arXiv:2104.07183*, 2021. [Online]. Available: <https://arxiv.org/abs/2104.07183>
- [10] M. Keshka, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, and A. Y. Zomaya, "An Explainable Deep Learning-Enabled Intrusion Detection Framework in IoT Networks," *Internet of Things*, vol. 21, 2023.
- [11] H. Benaddi, M. Jouhari, N. Laamech, A. Motii, and K. Ibrahim, "Lightweight Intrusion Detection in IoT via SHAP-Guided Feature Pruning and Knowledge-Distilled Kronecker Networks," *arXiv preprint arXiv:2512.19488*, 2025. [Online]. Available: <https://arxiv.org/abs/2512.19488>
- [12] A. N. Gummadi, J. C. Napier, and M. Abdallah, "XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems," *IEEE Access*, vol. 12, pp. 73451–73470, 2024.
- [13] L. C. Fourati, M. Awad, M. Ben Ali, and W. Jaafar, "Large Language Models for Cyberattack Defense: A Critical Survey," *Knowledge and Information Systems*, vol. 68, p. 113, 2026, doi: 10.1007/s10115-026-02736-y.
- [14] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, and D. Meng, "When LLMs Meet Cybersecurity: A Systematic Literature Review," *Cybersecurity*, vol. 8, article 55, 2025, doi: 10.1186/s42400-025-00361-w.
- [15] P. R. B. Houssel and S. Layeghy, "Towards Explainable Network Intrusion Detection using Large Language Models," *arXiv preprint arXiv:2408.04342*, 2024. [Online]. Available: <https://arxiv.org/abs/2408.04342>
- [16] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, T. Lestable, and N. S. Thandi, "Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices," *IEEE Access*, vol. 12, pp. 23733–23750, 2024, doi: 10.1109/ACCESS.2024.3363469.
- [17] M. Algarni, M. Y. Dahab, A. A. Alsulami, B. Alturki, B. M. Alghamdi, R. A. Alsemmeiri, and R. Alakhtar, "An Intrusion Detection System Based on Lightweight BERT with Meta-Classifer for Internet of Things Environments," *PeerJ Computer Science*, vol. 12, e3590, 2026, doi: 10.7717/peerj-cs.3590.
- [18] T. Ali and P. Kostakos, "HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)," *arXiv preprint arXiv:2309.16021*, 2023. [Online]. Available: <https://arxiv.org/abs/2309.16021>
- [19] P. R. B. Houssel, S. Layeghy, P. Singh, and M. Portmann, "eX-NIDS: A Framework for Explainable Network Intrusion Detection Leveraging Large Language Models," *Computers & Electrical Engineering*, vol. 129, Art. no. 110826, 2026, doi: 10.1016/j.compeleceng.2025.110826.
- [20] J. Kim, H. Maathuis, and D. Sent, "Human-Centered Evaluation of Explainable AI Applications: A Systematic Review," *Frontiers in Artificial Intelligence*, vol. 7, Art. no. 1456486, 2024.
- [21] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017.
- [22] A. Ferrag, L. Maglaras, H. Mosavi, et al., "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IIoT and IIoT Applications for Centralized and Federated Learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022.
- [23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and P. W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [24] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [25] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and Aggregating Feature-Based Model Explanations," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 3016–3022.