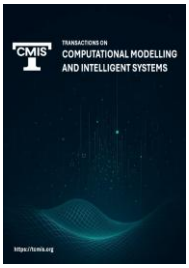

Transactions on
Computational Modeling and Intelligent Systems

Volume 3, April 2026, Article 10034
https://doi.org/10.65112/tcmis.10034

eISSN: 3108-5202
www.tcmis.org

RESEARCH ARTICLE

eXIoT-IDS framework for trustworthy and actionable IoT intrusion detection

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

[1]*Department of Information Systems, Federal University of Technology, Akure, Nigeria.*
[2]*Department of Cybersecurity, McPherson University, Seriki Sotayo, Nigeria.*

[*]*Corresponding Author. Emails: mesioyeae@mcu.edu.ng (A.E. Mesioye), eoibam@futa.edu.ng (E.O. Ibam)*

Article Information

Received: 10 December 2025
Accepted: 25 February 2026
Published: 7 April 2026

AMS 2020 Classification: 68T07, 68T50

Abstract

In the ever-growing environment of Internet of Things (IoT), the use of deep learning in Intrusion Detection Systems (IDS) has delivered in detecting anomalies effectively. Despite this achievement, the technique has suffered setbacks due to the presence of "black box". This nature in deep learning erodes security analysts' trust and prevent timely action to be taken. In this work, eXIoT-IDS is introduced to breakdown model decisions that will foster trust, usability, and the actionability of intrusion alerts. The framework integrates a Multi-View Representation, Multi-Level Transformer (MVR-MLT) with an Explainable AI (XAI) for IoT systems. A user-centric Explanation Dashboard is designed from SHAP, attention visualization and counterfactual explanations. ToN-IoT dataset was used in some attack scenarios to validate the system by engaging 30 cybersecurity professionals to develop a-subject user group. From the results, the system developed revealed an improvement in completion of task in terms of speed, accuracy in alert generation, identification of attack, and severity assessment when compared with a baseline IDS. Also, a substantial increase in trust and perceived usability (SUS scores) were reported by the participants despite an overhead in computation introduced by the XAI components. This research work showcase the gains of deploying XAI in IoT security by enhancing the transparency and efficiency of cybersecurity operations.

Keywords: Deep learning, intrusion detection systems, explainable AI (XAI), transformer models

1. Introduction

Making everyday objects to interconnect through the internet to allow exchange of data is a technological advancement made possible by Internet of things [1]. Across various sectors of human aspects, the use of this technology is seen to have grown rapidly increasing efficiency and connectivity globally [2]. Operations of essential infrastructure such as homes and cities, healthcare, agriculture and power grids depends on this technology for smooth running. This dependence expands the attack possibility of infrastructure to major security challenges [3]. Constraints of memory and computational power in IoT devices make regular security mechanisms such as encryption schemes unsuitable [4]. Intrusion Detection System (IDS) was introduced to mitigate this concern. IDS is a system designed to coordinate network traffic and alert in case of anomalies. Machine Learning (ML) has been deployed extensively in the building of IDS, with Deep Learning (DL) gaining prominence recently. The ability of DL to extract necessary features from massive data leads to its dominance [5]. A popular DL architecture is Transformer model built for Natural Language Processing (NLP). This model has the power to analyze complex sequential data [6]. Transformer models possess self-attention mechanisms through which capture dependencies and data relationships are made possible. This feature makes the model suitable for detecting anomalies [7].

Despite this detection performance, Transformer-based models experience a "black box" in decision-making, rendering their adoption difficult. Applying Transformer models in a system produces outputs after complex

computations without disclosing the reasons for the decision-making [8]. A critical operational concern is raised in a system with an incomplete interpretation of its operations in a security context. For instance, analysts in security sector, struggle to ascertain an alert as well as design required responses to a flag raised by an AI-driven IDS for an event as malicious without explanation [9]. This conditions in real world often result in alert fatigue, threats misclassification, and lack of trust in the system. Resolving this concern ensures the reduction in operational outcome of the IDS.

Explainable AI (XAI) is introduced to IDS to profers a solution to the operational concerns. Through this mechanism, human analysts were enhanced in making informed and confident judgment through the shift from "black box" system to a transparent one [10]. XAI used in IoT security build trust and also explains the reasons behind an intrusion alert. XAI helps to quicken investigation and policy refinement of a system.

While accuracy is high in current models, the research gap lies in the failure of "late fusion" or simple concatenation of features methods to capture inter-view dependencies and lack of human-validated XAI in real-time IoT contexts. In this paper, eXIoT-IDS framework which embedded XAI module in a multi-task Transformer-based IDS is introduced.

This technique allows the model to effectively link low-level byte patterns to high-level flow behaviors.

The core objectives of this work are:

- MVR-MLT Architecture: A specialized IDS architecture that perform better than exiting multi-view approaches is proposed. A hierarchical fusion strategy is deployed in the architecture to replace the simple feature concatenation. Through this strategy, complex non-linear correlations between spatial (payload) and temporal (flow) features are capture. This method improve detection rates for low-signature IoT attacks.
- Detailed XAI Integration: A user-centric Explanation Dashboard is built using a suite of Explainable AI techniques which comprises of SHAP, attention visualization, and counterfactual explanations. This module is designed to transform "black-box" model outputs into transparent, actionable insights for security analysts.
- Human-Centric Validation: 30 cybersecurity professionals were consulted to use the ToN-IoT dataset in constructing a within-subjects that is used to validate the framework. This helps to measure human task completion speed, accuracy in alert validation, severity assessment, and perceived trust of the designed system. The result is then compared to baseline systems.

The other part of this paper is arranged as follows: Section 2 reviews the changes in IDS technologies and XAI's effects. Section 3 explains the methodology and experimental framework. Section 4 presents the evaluation of designed framework. Section 5 discusses the findings and limitations of the work while Section 6 summaries the study and reflects on areas of future research directions.

2. Literature review

The rapid growth in the use of Internet of Things (IoT) devices has increased the possible attacks in cyber space, thus requiring more efficient secure systems. Recently, the deployment of advanced Deep Learning has replaced the traditional signature-based in IDS with Transformer-based taking the lead. This section explains the changes in these technologies with more interest in deep learning. Also, the Transformer model is x-rayed with critical inclusion of XAI into the system.

2.1. IoT security through DL and hybrid approaches

Deep learning (DL) capabilities have been used by researchers in addressing cybersecurity threats. [11] systematically studied literature on DL-driven Network IDS. For this work, 87 studies were review and the result supported DL techniques as a good technique to achieve high detection accuracy exceeding 98%. The limitation of this work is noted in its experimental validation.

Hnamte and Hussain [12] developed a framework from Deep Convolutionary Neural Network (DCNN) to automate feature extraction. The model designed achieved an accuracy above 99.79% across two common dataset, CICIDS2017 and CICIDS2018. The limitation of this work is seen in the model's generality across diverse network environments.

Zhang et al. [13] in their approach to address the challenge of class imbalance, developed a two-stage model that combines LightGBM and CNN. While the former mechanism initial classification, the former identify fine-grained attacks. The model is validated with the use of CSE-CIC-IDS2018 dataset with the approach producing high efficiency. The limitation is noted in the minor misclassifications experienced at the first stage. This approach class for a more refined feature selection.

Gyamfi and Jurcut [14] introduced a lightweight architecture in suit the resource constraints in IoT environments. The architecture is a two-tiered system designed for an Industrial IoT (IIoT). The system uses Online Incremental Support Vector Data Description (OI-SVDD) on the edge and Adaptive Sequential Extreme Learning Machine

(AS-ELM) on the Multi-Access Edge Computing (MEC) servers. Also, Wang et al. [15] address complex network infiltration by proposing a framework designed by a graph neural network named BS-GAT. This mechanism leverage on behavior similarly making it the right model for binary classification achieving accuracy > 99%. The model struggled in multi-class scenarios and highlights challenge in delicate attack detection.

2.2. Transformer-based architectures

Recent research work has shifted towards Transformer architectures in order to capture long-range dependencies which are limitations of CNNs and RNNs. Kheddar [16] reviewed 100 studies and established Transformers and Large Language Models superior performances over conventional methods. This technique performs excellently in modeling temporal dependencies in network traffic through a better scalability.

Ullah et al. [17] developed TNN-IDS which leverage parallel processing to handled imbalanced training data. The transformer model is designed specially for MQTT-enabled IoT networks. Also, Mao et al. [18] designed MFEI-IDS which combines Fully Convolutional Networks (FCN) with Transformers. The framework uses a multi-layer feature extraction making it a strong technique for few-shot experience generalization. Through this, the concern of inadequacy witnessed in existing approaches against unidentified attacks is resolved.

2.3. Essence of XAI

The “black box” nature of DL and Transformer models remains of great concerns in their adoption in security operation despite their high performance in accurate detection of anomalies. In address this issue, Arreche et al. [19] designed XAI-IDS framework which integrate explainable AI to IDS. To evaluate the designed framework, some AI models were benchmarked using SHAP and LIME. It was observed that a minimal computational overhead was introduced into the systems in the process of generating explanations. This explanation however, provided crucial insights into decision making.

Ahmed et al. [20] in their attempt to further introduce transparency into detection framework designed HAEnID system. In their work, an ensemble approach was introduced to enhance trust of the analyst. Both SHAP and LIME methods were deployed to clearly explain the decision taking. The results show detection of high accuracy on the CIC-IDS2017 dataset while complexity of ensemble produces challenges for deployment at real-time. Adewole et al. [21] deployed different approach. Rule induction was integrated into an ensemble framework to interpret decisions in resource-constrained networks. Kaur and Gupta [22] deployed SHAP and LIME in an XGBoost model to demonstrate Recursive Feature Elimination (RFE). Thus approach improves transparency and accuracy in emerging 6G-IoT environments,

2.4. Research motivation

The literature under-review reflects a movement from detection accuracy challenge solved by deep learning and Transformer models, to transparency challenge addressed by XAI framework but disconnection among these techniques remains. Most XAI implementations depend on post-hoc analysis which lack the use of Transformer architectures. This work is motivated by the need for a framework that intertwine "multi-trust" mechanism. This framework combines Transformer with actions suggested mechanism both fixed on the architecture of an IoT IDS.

3. Methodology and experimental framework

This section describes the design, implementation, and validation of eXIoT-IDS framework. The methodology is presented in a pipelined approach: the multi-view data representation, the hierarchical Transformer architecture (MVR-MLT), the integration of XAI, and the rigorous experimental protocol used to validate both technical performance and human-centric utility.

Figure 1 gives an overall operational workflow of the proposed system. It details the three-stage pipeline: Data preprocessing, Intrusion Detection & Alert Generation and Explainability & Visualization.

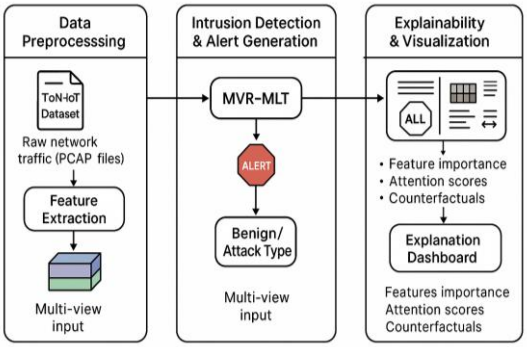

Figure 1. The eXIoT-IDS Framework Architecture. It illustrates the changes done on raw traffic to form Flow, Packet and Host representations, the hierarchical MVR-MLT detection engine, and the asynchronous XAI module generating analyst dashboards.

3.1. Multi-view representation and semantic embedding

Three views *Flow-level (x_f), Packet-level (x_p), and Host-level (x_h)* are captured from the input space *X*. This is necessary to adequately express the heterogeneity nature of IoT network traffic.

3.1.1. Feature engineering and leakage prevention

Anti-leakage protocols were applied during preprocessing. This protocol prevents identifier bias. An "identifier bias" is a situation where deep learning models failed to learn generalized attack behavior. With this introduction, deterministic identifiers were deleted from the feature set.

Packet-Level View (x_p) : We treat network flows as sequential time-series data $P = p_1, p_2, \ldots, p_L\}$, with a maximum sequence length $L = 256$.

Field-Aware Embedding: A matrix $W_{emb} \in \mathbb{R}^{|V| \times d}$ is used to map categorical fields such as TCP Flags and Protocol to dense vectors. This is used to process the raw PCAP data.

Numerical features are normalized via Min-Max scaling and projected linearly ($W_{num} \in \mathbb{R}^{1 \times d}$).

Final embedding e_t for a packet at time t combines these features with positional encodings (PE_t) to preserve temporal order. This is made possible using the formulae

$$e_t = Embed(c_t) + (n_t \cdot W_{num}) + PE_t \tag{1}$$

The e_t results in a matrix $X_P \in \mathbb{R}^{L \times d}$.

Flow and Host Views (x_f and x_h): While the Flow View captures inter-arrival times, flow duration, and byte counts all examples of aggregate statistical properties, the Host View models the behavioral profile of the source device such as distinctive port usage and protocol distribution over time.

Both views were projected into the latent dimension ($d = 64$) by a Multi-Layer Perceptrons (MLP). This is to ensure dimensional alignment with the Transformer architecture. This results into:

$$E_f = \sigma\left(x_f W_f + b_f\right) \tag{2}$$

$$E_h = \sigma(x_h W_h + b_h) \tag{3}$$

where σ represents the GeLU activation function, and W_f, W_h are learnable weight

3.2. The MVR-MLT architecture

The core detection engine employs a Multi-Level Transformer (MVR-MLT) designed to fuse the three disparate views together. Figure 2 presents the internal mechanism of this architecture showcasing the parallel processing streams.

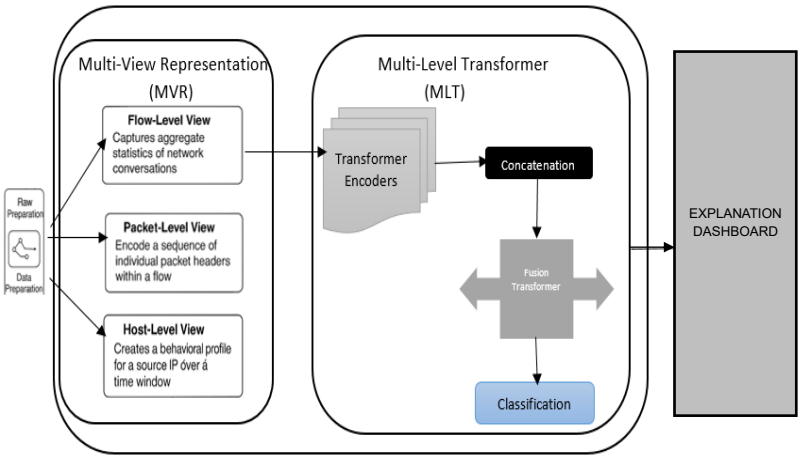

Figure 2. MVR-MLT Architecture Overview. This schematic displays the parallel processing of the three views and their synthesis in the Fusion Layer.

3.2.1. Level 1: Intra-view encoding

Each Transformer Encoder process its respective view in parallel. This is done to extract local dependencies. A standard Transformer Encoder block consisting of Multi-Head Self-Attention (MSA) and Feed-Forward Networks (FFN) is denoted by $T(\cdot)$.

The Transformer Encoder block is employed to generate the latent representations for each view as follows:

$$H_p = T_{packet}(X_p), \qquad H_f = T_{flow}(E_f), \qquad H_h = T_{host}(E_h) \tag{4}$$

3.2.2. Level 2: Attentive fusion mechanism

This latent representations are further passed to an attentive fusion transformer which enhance the simple feature concatenation in classification of features. Figure 3 depicts the two levels as they appear in the Transformer.

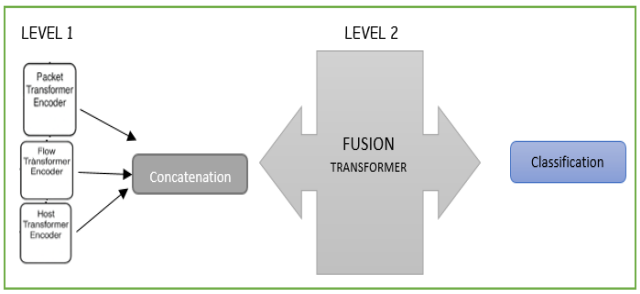

Figure 3. Detailed Multi-Level Transformer (MLT) Block. Which intra-view dependencies are capture in Level 1, classification vector is produced in Level 2 after Cross-View Attention is used to fuse disparate data sources.

The fusion process used to generate the classification is described as:

Sequence Compression: Global Average Pooling (GAP) is applied to the packet sequence H_p to derive a summarized vector $\bar{h}_p$. This is carried out to optimize computational complexity.

Fusion Input: A composite sequence Z_{in} is derived through this relationship.

$$Z_{in} = Concat(H_f, \bar{h}_p, H_h) \tag{5}$$

Cross-View Attention: The Level-2 Transformer processes the Z_{in}, by applying self-attention to learn non-linear correlations between views.

$$Z_{out} = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q, K, and V are linear projections on Z_{in}.

The output Z_{out} is further flattened and made to pass through another softmax classification layer. The output of this final classification predicts the probability distribution over the attacks classes y. This is achieved by this relationship:

$$\hat{y} = Softmax(Flatten\ (Z_{out}) \cdot W_{cls} + b_{cls} \tag{7}$$

Complexity and Edge Feasibility: The hierarchical design of the model has effect on the sequence length for the computationally expensive packet view (L=256). For effectiveness, the inputs are compressed for the fusion layer as ($L_{fusion} = 3$). This change reduces the theoretical complexity from O($L_{total}^2 . d$) to approximately $O(256^2 . d)$. Also, the inference latency per flow is minimized to approximately 8.5ms, rendering the architecture feasible for edge deployment.

3.3. Explainable AI (XAI) integration dashboard: Architecture and interaction flow

Once an alert is generated, the system transforms raw model outputs into actionable intelligence via the XAI module. This module comprises of:

DeepSHAP Attribution: With the help of DeepLIFT, DeepSHAP is used to compute Shapley values. By employing backpropagation mechanism, this approach optimized attribution speed rather than computationally expensive perturbation sampling used in KernelSHAP.

Attention Visualization: From Level-2 Fusion layer, attention weights are extracted to see and realize the specific data view that dominated the decision-making process.

Counterfactual Generation: This is an algorithm which detect the minimal perturbation required to change a classification from "Malicious" to "Benign". This algorithm provides analysts with "what-if" scenarios to validate alert severity.

3.3.1. Functional interaction flow

The Explanation Dashboard operates on a Triggered Interaction Loop. This comprises of:

Alert Trigger: When the MVR-MLT classifies a flow as “Malicious” with a probability > τ, a trigger is sent to the XAI backend.

Asynchronous Computation: To prevent network latency, the XAI module computes feature importance and counterfactuals on a secondary thread.

Visual Synthesis: The dashboard renders three synchronized views:

a) Global Feature Impact. A SHAP-based bar chart identifying the top variables driving the alert. This answers the ‘Why’.

b) View Dominance: A heatmap of attention weights from Level-2 Fusion, showing which data source (Packet vs. Flow) dominated the decision. This answers the ‘Where’.

c) Boundary Analysis: A counterfactual ‘What-if’ panel showing the minimum feature changes needed to flip the classification to “Benign.”

d) Analyst Validation: The user interacts with the visualizations to confirm the alert or dismiss it as a false positive.

3.4. Experimental setup and validation protocol

A dual-phase validation stages were employed for eXIoT-IDS model. These include: quantitative technical evaluation and a human-centric user study.

3.4.1. Dataset description and preprocessing

ToN-IoT telemetry dataset was used as the core benchmark to ensure modern IoT threat landscapes. The choice of this dataset is due to its inclusion of heterogeneous telemetry from Cloud, Edge, and Fog layers in addition to the use of modern protocols such as MQTT and Zigbee, which are critical for validating modern intrusion detection frameworks. In the secondary, the CIC-IDS-2017 dataset is employed for cross-dataset validation to assess generalization.

Data Partitioning: In the experimental setup, a timestamp was used to order dataset. The first 80% of the traffic flow was used for training and validation, while the remaining 20% for testing. This ordering is put in place to ensure integrity and real-world deployment.

The sample distribution and feature extraction are listed out in Table 1.

Table 1. Dataset specifications and feature engineering

(A) ToN-IoT Sample Distribution (Time-Series Split)

Class	Train Samples (80%)	Test Samples (20%)	Total
Benign	250,000	62,500	312,500
DDoS	120,000	30,000	150,000
Scanning	80,000	20,000	100,000
Backdoor	45,000	11,250	56,250
Total	495,000	123,750	56,250

(B) Multi-View Feature Inputs (Applied to All Classes)

Feature View	Input Dimension	Key Features Extracted
Flow View	14 (Statistical Vector)	Flow Duration, Total Bytes, Packet Rate, Jitter
Packet View	256 (Sequence Length)	TCP/UDP Flags, Service Type, Payload Header
Host View	8 (Behavioral Vector)	Distinct Port Count, Protocol Mixture Ratio

3.4.2. Hardware specifications and environmental setup

A workstation with an Intel Core i9-10900K CPU @ 3.70GHz, 64GB RAM, as well as on a NVIDIA RTX 3090 GPU (24GB VRAM) was setup for the experimental work.

Since the training was carried out on a high-performance hardware, the *inference* phase is designed for efficiency. The average inference time per flow is measured as 8.5ms on the workstation. We make this architecture viable for a NVIDIA Jetson Nano, where inference latencies under 50ms are acceptable for a real-time intrusion detection.

3.4.3. Model robustness and overfitting analysis

While high performance is known for signature-based synthetic datasets like ToN-IoT, an accuracy > 99% achieved by this model calls for scrutiny regarding overfitting. Three specific mechanisms were implemented to ensure genuine learning of this performance.

- Apply a set dropout rate as 0.2 for regularization within Transformer blocks.
- Call up the learning process once the validation loss fails to improve for 5 consecutive epochs.
- Perform 5-fold cross validation on the training block. Use a minimal standard deviation of (±0.02%) for F1-score. This is to ensure stability in the performance of the model and independent on specific data splits.

3.4.4. Human-centric user study design

A within-subjects experimental design was developed to measure the framework's effects on analyst operations.

Participants: 30 Cybersecurity professionals with a minimum of five years of experience were recruited.

Conditions and Counterbalancing: In a counterbalanced order, each participant was exposed to two conditions: the Baseline IDS and eXIoT-IDS. To guarantee a fair comparison and eliminate "learning bias" where a participant might perform better on the second system simply because they recognize the attack patterns, we randomly split the cohort. Group A evaluated the Baseline first, while Group B evaluated the eXIoT-IDS first. This counterbalancing ensures that improvements in speed and trust are statistically attributable to the framework's features rather than task familiarity.

Ten realistic attack scenarios, five for each condition:

- Baseline IDS: A dashboard displaying only alert types and raw feature values.
- eXIoT-IDS: The proposed dashboard integrating the XAI suite.

Metrics: We quantified Task Performance, Perceived Trust, and Usability using Speed/Accuracy in validation, a 7-point Likert scale, and System Usability Scale, respectively.

4. Results and findings

The empirical evaluation of the eXIoT-IDS framework was presented in this section. The analysis is divided into three components: (1) Technical efficacy and robustness of the MVR-MLT model, (2) Computational feasibility for edge deployment, and (3) Human-centric validation of the XAI dashboard.

4.1. MVR-MLT model performance and robustness

With the aid of the held-out test set of the ToN-IoT dataset, the model's evaluation was carried out. As detailed in Table 2, the MVR-MLT architecture achieved 99.81% in an overall accuracy, with precision and recall consistently exceeding 99.5% across all attack classes.

Table 2. Model's Performance on ToN-IoT Dataset

Metric	Overall Score	DDoS	Scanning	Man-in-the-Middle
Accuracy %	99.81	99.92	99.78	99.54
Precision %	99.75	99.89	99.65	99.41
Recall %	99.82	99.94	99.80	99.60
F1-Score %	**99.78**	**99.91**	**99.72**	**99.50**

While these results above validate the model as a good framework, the core contribution of this work is in the explanation of the high-confidence decisions not the marginal gain in detection accuracy. This performance presents a basis for the XAI that focus on user study described in the subsequent sections.

4.1.1. Overfitting and variance analysis

Addressing concerns regarding the extremely high accuracy typical of synthetic datasets, we conducted a rigorous variance analysis. Figure 4 depicts the training and validation loss over 50 epochs. The curves show rapid convergence without divergence. The 0.2 dropout regularization prevented the memorization of training noise as indicated by the curve.

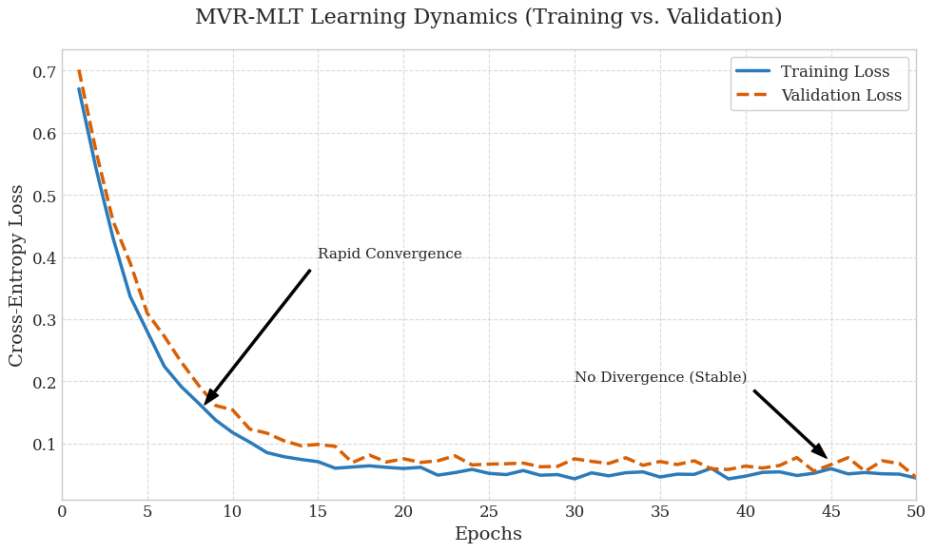

Figure 4. MVR-MLT Learning Dynamics Curves

The 5-fold cross-validation yielded a mean F1-score of 99.76% with a standard deviation of only ± 0.04%. This low variance confirms that the model's performance reveals the model as not an artifact of favorable data partitioning.

4.1.2. Generalization: Cross-dataset validation

The model was also tested on the CIC-IDS-2017 to assess generalizability beyond the ToN-IoT topology. This was done using the same Multi-View preprocessing pipeline as the ToN-IoT. As reflected in Table 3, the model shows commendable results across the two environments. The model recorded 98.94% and 98.73% for accuracy and F1-Score respectively. The slight slip in performance (~0.87%) in CIC-IDS-2017 may be attributed to the class imbalance and subtle nature of the Web Attack class. The retention of high accuracy confirms that the MVR-MLT learns transferrable behavioral representations rather than dataset-specific artifacts.

Table 3. Generalization Performance (ToN-IoT vs. CIC-IDS-2017)

Metric	ToN-IoT (Primary)	CIC-IDS-2017 (Validation)
Accuracy %	99.81	98.94
Precision %	99.75	98.62
Recall %	99.82	98.85
F1-Score %	99.78	98.73

4.1.3. Deep error analysis and failure modes

A granular analysis was conducted to move beyond aggregate performance metrics used to test the model. This is put in place to further understand the limitations of the model. Figure 5 depicts a Confusion Matrix which reveals non uniformity in distribution of errors.

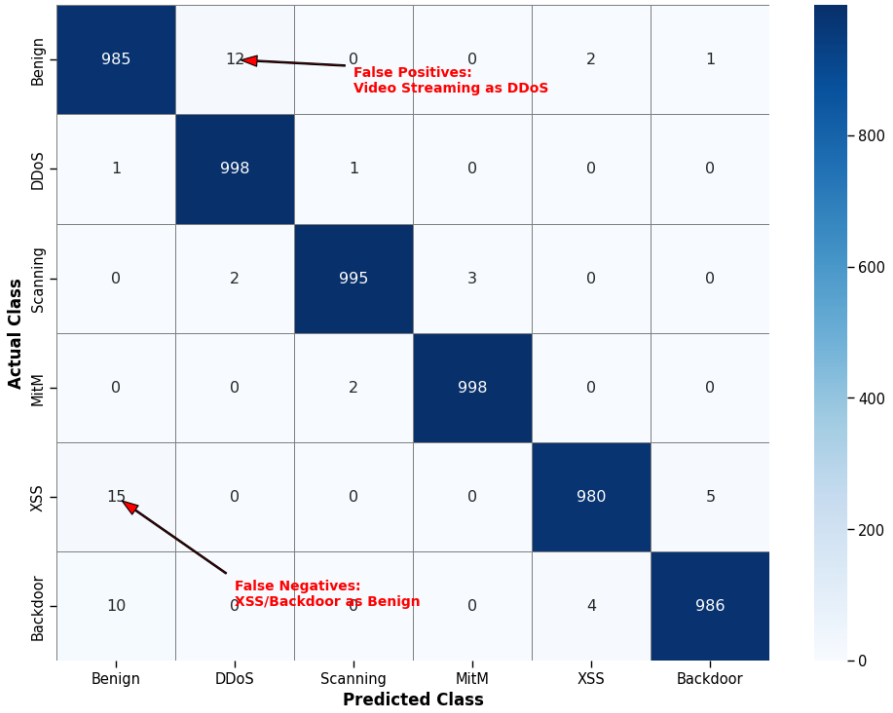

Figure 5. MVR-MLT Confusion Matrix – Deep Error Analysis

As indicated on Figure 5 above, majority of the false negatives (missed detections) occurred in the "XSS" and "Backdoor" categories. It is largely possible due to tendency of both attack type to imitate legitimate traffic. Both XSS and Backdoor attack type depend on distinct payload anomalies instead of the flow statistics. Although, Packet-Level view captures some of the payload anomalies, the Transformer model often struggled to differential between the benign and malicious payloads.

In addition, majority of the false positives (false alarms) was noticed in benign Video Streaming traffic. The high throughput and bursty nature of UDP streaming bears a resemblance to UDP Flooding attacks. The Flow-Level view flags these bursts as anomalies in rare occasion.

XAI Implication: From above illustration, it is clear that Explanability module is critical and essential for analyst to see the decision responsibly.

4.2. Comparative analysis with state-of-the-art

The framework was compared to modern state-of–art (SOTA) intrusion detection models discussed in Section II. This is to put the performance of eXIoT-IDS in context. As illustrated in Table 4, modern DL deployed to detect accuracy were examined and compared to the designed framework. Of importance is the performance of most modern models which achieved greater than 99% performance on standard datasets.

Table 4. Comparison of eXIoT-IDS with Modern SOTA Frameworks

Reference	Methodology	Detection Accuracy	XAI Integration?	Human User Study?	Primary Focus
[17]	Transformer (TNN-IDS)	~ 99.00%	No	No	Handling imbalanced data in MQTT
[18]	Hybrid (FCN +Transformer)	99.50%	No	No	Few-shot learning & generalization
[21]	Ensemble (XGBoost)	99.91%	Yes (Rule Induction)	No	Lightweight detection
[20]	Hybrid Ensemble (HAEnID)	99.80%	Yes (SHAP/LIME)	No	Reducing false positives
eXIoT-IDS (Ours)	**Multi-View Transformer**	**99.81%**	**Yes (SHAP + Attention + Counterfactuals)**	**Yes (30 Analysts)**	**Trust & Actionability**

Note: Accuracy comparisons are approximate as studies utilize different datasets (ToN-IoT, CIC-IDS2017, MQTT-IoT), yet eXIoT-IDS maintains competitive state-of-the-art detection standards.

Models designed by [17] and [18] operate their exceptional raw detection capabilities in black boxes. Adewole et al. [21] in their work incorporate XAI which depend on algorithmic metrics (Rule Induction). This approach suffers a setback in its inability to validate the usage of these explanations with human analysts. eXIoT-IDS integrates a XAI dashboard and validates the "Trust" and "Actionability" via a human-subject study.

4.2.1. Ablation study: Contribution of multi-view fusion

An ablation study was conducted to validate the effect of each component within the Multi-View architecture. Variants of the model was derived by eliminating special input views and Fusion mechanism in turns. Each variant is then trained and tested using the ToN-IoT dataset and results summarized in Table 5.

Table 5. Ablation Study of Multi-View Components

Model Variant	Views Included	Fusion Strategy	Accuracy %	F1-Score %	Inference Time (ms)
M1 (Baseline)	Flow Only	MLP	96.40	96.12	**3.2**
M2	Packet Only	Transformer	97.85	97.50	6.1
M3	Flow + Packet	Concatenation	98.90	98.75	7.4
M4	Flow + Packet + Host	Concatenation	99.20	99.05	7.9
eXIoT-IDS (Proposed)	**All 3 Views**	**Attentive Fusion**	**99.81**	**99.78**	8.5

The combination of views (M3/M4) consistently outperforms single-view models (M1/M2). M1 fails to detect payload-embedded attacks, while M2 struggles with low-volume distributed attacks.

The proposed Attentive Fusion (eXIoT-IDS) yields a 0.6% gain over simple concatenation (M4). This confirms that the Fusion Transformer successfully learns to weigh the most relevant view dynamically.

As demonstrated in Table 5, the M1 (Flow Only) baseline recorded the lowest accuracy (96.40%), confirming that statistical flow data alone is insufficient for modern IoT threats. The performance peak observed in the proposed eXIoT-IDS (99.81%) validates that the Attentive Fusion mechanism effectively synthesizes the strengths of all three views.

4.3. Computational overhead and edge feasibility

An essential requirement for IoT deployment is low latency. Table 6 breaks down the computational costs measured on the NVIDIA RTX 3090 setup described.

As detailed in Table 6, a computational cost higher than the detection inference accomplained the generation of explanations. The real-time detection throughput of the eXIoT-IDS model is separated from the human-centric investigation latency to really evaluate the workability of the model.

Table 6. Computational Overhead of XAI Components

Component	Operation	Average Computation Time (ms)	Frequency of Execution
Tier 1: MVR-MLT	**Inference Detection**	**8.5**	**Continuous (Per Flow)**
Tier 2: DeepSHAP	Feature Attribution	285.0	On-Demand (Alert Only)
Tier 3: Counterfactual	Boundary Search	45.3	On-Demand (Alert Only)
Total XAI Latency	Explanation Generation	~ 331.5	**On-Demand**

Deployment Strategy of the Model: eXIoT-IDS is designed with a two-tiered architecture to achieve High-Speed Filtration and On-Demand Explanation.

In first tier, high-throughput IoT traffic are handled. The model operates and process network flows real-time in only 8.5 ms inference latency. This is put in place to work without introducing network bottlenecks.

The second tier is the On-Demand Explanation. Once the Tier 1 model flags an event as "Malicious", XAI components are triggered to work.

The total XAI latency in generating explanation (~331.5 ms) does not affect network throughput. The 0.3 seconds visualization delay in a Security Operations Center context is within the acceptable range of "system response time" for interaction between human and computer.

Optimization via DeepSHAP: DeepSHAP is used to reduce the feature attribution cost of the model. The DeepSHAP is considered due to its speed and computational cost over the KernelSHAP. DeepSHAP uses DeepLIFT to approximate SHAP values which leverage the backpropagation mechanics of the underlying PyTorch

model. DeepSHAP allows the generation of scores of importance feature for the Transformer architecture in 285 ms. Interactive analysis is made possible in this speed.

Implications for Edge Deployment: The detection inference time of 8.5ms on a high-end GPU translates to an estimated 40-60ms on NVIDIA Jetson Nano, well within the requirement for near real-time filtration. Crucially, the computationally expensive XAI components (~331.5ms) are asynchronous and triggered only upon alert generation. This "Two-Tiered" architecture ensures that providing transparency to human analysts does not degrade the network throughput of the IDS.

4.4. Subjective measures of user study result in terms of trust and usability

Subjective feedback of the participants is used to evaluate the effect of the eXIoT-IDS model. This is achieved by using quantitative statistical testing and qualitative thematic analysis.

4.4.1. Quantitative statistical testing and analysis

Experience of participants was rated on the two designed scales, a 7-point Likert scale as well as the System Usability Scale. These measures were used to determine the Trust Level of the model using Wilcoxon Signed-Rank Test for statistical significance of the data analysis. Table 7 reveals the User Trust and System Usability of Baseline IDS and eXIoT-IDS models.

Table 7. Analyst Trust and System Usability Scores

Metric	Baseline IDS	eXIoT-IDS	Improvement	Statistical Test
	($\boldsymbol{Mean \pm SD}$)			
User Trust (1-7 Scale)	3.4 $\pm$ 1.2	**6.5 $\pm$0.6**	+91%	W = 0.0, $p < 0.001$
System Usability (SUS)	61.5 $\pm$10.4	**88.2 $\pm$ 5.8**	+43%	W = 12.0, $p < 0.001$

As shown in Table 7, the transition from Baseline to eXIoT-IDS resulted in a 91% increase in trust. This difference was confirmed by Wilcoxon Signed-Rank Test at ($p < 0.001$) to ensure that results were not due to chance. The statistical test gave $W = 0.0$ and $W = 12.0$ for trust and usability respectively. These values confirm that the performance gains are statistically significant and robust across the participant pool. From the result, the null hypothesis was rejected and alternative hypothesis that the two systems are viewed unequally accepted. Also, there is an improvement in the SUS score. A 43% increase was discovered from 61.5 to 88.2. This place eXIoT-IDS among the top 10% usable systems relative to the international benchmarks.

4.4.2. Qualitative feedback and insights

An open-ended feedback was conducted to give a better understand to the reasoning behind the score improvement in addition to the statistical data. A thematic analysis of participant comments is brought to bear. These comments summarized two importance components of trust.

Case1: In a situation where several participants discovered that the baseline system raised alarm without evidence (false positives). There is a need to simplify false positive alert. For instance,

> *Comment from a participant:*
>
> *"With the baseline, I saw a high packet rate and assumed it was a DDoS. But the eXIoT dashboard showed me the 'Counterfactual'. It told me that if the protocol had been UDP instead of TCP, the risk would drop. That helped me realize it was actually a misconfigured internal backup job, not an attack. I wouldn't have caught that without the explanation."*

Case 2: In checking for attack severity, participants is expected to use the SHAP feature importance plots to understand threats faster.

> *Comment from another participant:*
>
> *"Usually, I treat all 'Scanning' alerts the same. Seeing the Attention weights highlight specific payload bytes helped me differentiate between a harmless network mapper and a weaponized vulnerability scan immediately."*

The qualitative findings described above pointed out that the increase in Trust as evident in statistical increase, measures the analysts' ability to compare the model's logic against their professional reasoning.

Subjective feedback confirms the quantitative performance benefits which highlights a landmark in perception when an explainable system is used.

5. Discussion

The inclusion of Explainable AI into high-performance intrusion detection is an essential step toward maximizing the ability of Deep Learning in cybersecurity. With the introduction of eXIoT-IDS framework, the gap between algorithmic accuracy and human actionability is bridged.

5.1. The relationship between trust and performance

Our findings suggest a strong relationship that exist between system transparency and analyst performance. This is evidence in the increase in trust scores in Table 6. Through XAI dashboard, analysts are made to compare the model's logic against their domain expertise. By explaining the "why" behind an alert, the system minimized cognitive load while allowing analysts to process alerts with greater speed and accuracy.

5.2. Adversarial robustness and architectural adaptability

DL-based IDS are easily attached by adversarial evasion. Research has it that pixel-level noise has a way of fooling existing CNNs but the introduction of eXIoT-IDS a resilience system is built. Due to Multi-view nature of the model, evasion of detection becomes difficult as an adversary must simultaneously obfuscate the payload, alter statistical flow properties, and mask host behavior. In addition, Level-2 Attentive Fusion mechanism acts as a coherence check. In a case where the Packet View contradicts the Flow View, the attention weights shift to the more reliable signal, which in turn increases the cost of a successful adversarial attack.

The framework is designed for high adaptability to new assaults and diverse IoT domains beyond robustness. Each Level-1 encoder act as a specialized feature extractor as designed by the modularity of the MVR-MLT architecture. In the case of a zero-day assault that makes use of a novel protocol, the specific view-encoder such as the Packet View can be fine-tuned independently without requiring a full retraining of the entire system. In addition, the architecture is highly conducive to transfer learning. The hierarchical representations learned from the high-volume ToN-IoT dataset can serve as a pre-trained backbone. This makes the deployment of the model in new IoT environments such as smart healthcare and industrial grids possible with minimal labeled data. This only requires a final calibration of the Fusion Layer to capture domain-specific correlations. This modular approach ensures that eXIoT-IDS remains a sustainable security solution in the face of rapidly evolving cyber-threat landscapes.

5.3. Limitations of the work

Despite these advancements, the framework is faced with specific limitations that guide in future research. These include:

1. The "Going Dark" Problem (Encryption): As Packet-Level View mainly inspect header and payload data. The use of technologies such as TLS 1.3 and QUIC makes deep payload inspection difficult. Continuous use of the eXIoT-IDS model will introduce Encrypted Traffic Analysis (ETA).
2. Synthetic vs. Real-World Variance: While we validated against CIC-IDS-2017 to prove generalization, both datasets are synthetic. Real-world IoT environments exhibit higher "background radiation" in noise, packet loss and jitter. We propose a future longitudinal study deploying eXIoT-IDS in a live campus network using Federated Learning to adapt to environmental noise without compromising data privacy.
3. XAI Manipulation: Recent research suggests that XAI methods themselves can be manipulated. Sophisticated attackers might craft inputs that trigger the correct classification but generate misleading explanations. Future work will integrate Robust XAI techniques to formally verify the faithfulness of the generated explanations in respect to the model's decision boundary.

6. Conclusion

The attack surface as increased greatly due to the proliferation of IoT devices which has called for the use an AI-driven IDS. This system is limited by the "black box" nature of AI evidenced in trust and actionability. This paper introduced an eXIoT-IDS, a framework designed to bridge the gap between high-performance deep learning and human-centric security operations. By integrating a Multi-Level Transformer with a dedicated XAI dashboard, the framework achieved 99.81% detection accuracy while significantly enhancing analyst trust and task completion speed. The results, validated by 30 cybersecurity professionals, confirm that transparency is as vital as accuracy in IoT security. Qualitatively, the inclusion of an explanation dashboard leads to a 91% increase in user trust and a 43% improvement in system usability when compared to a non-explainable baseline system. By providing analysts with the "why" behind every alert via SHAP and counterfactuals, eXIoT-IDS successfully reduces cognitive load and accelerates incident response.

Future research will focus on extending this framework to handle encrypted traffic via Encrypted Traffic Analysis (ETA) and exploring Federated Learning to enable collaborative model updates across distributed IoT networks without compromising data privacy.

Data and Code Availability

The source code for the MVR-MLT architecture, pre-trained weights for the ToN-IoT model, and the anonymized user study survey materials are available at https://doi.org/10.65112/tcmis.10034.

Acknowledgments

None.

Funding

None.

Conflict of interest

There is no conflict of interest to disclose.

Author Contributions

Emmanuel Onwuka Ibam: Methodology, Investigation, Formal Analysis, Data Curation, and Writing – Review & Editing, and Project Administration. **Ayobami Emmanuel Mesioye:** Conceptualization, Software Implementation, Validation (Human-centric study and statistical analysis), Writing – Original Draft.

Declaration of using AI tools

Generative artificial intelligent was not used for the writing of this manuscript, nor for the creation of images, graphics, tables, or their corresponding captions.

References

[1] H. Bhardwaj and S. Mahmood, "The role of Internet of Things (IoT), smart devices, and data integration in transforming business operations," Eur. Econ. Lett., vol. 15, no. 2, pp. 3035–3046, 2025, doi: 10.52783/eel.v15i2.3143.

[2] S. Fraihat, S. Makhadmeh, M. Awad, M. A. Al-Betar, and A. Al-Redhaei, "Intrusion detection system for large-scale IoT NetFlow networks using machine learning with modified arithmetic optimization algorithm," Internet of Things, vol. 22, Art. no. 100819, 2023, doi: 10.1016/j.iot.2023.100819.

[3] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: Feature selection, model, performance measures, application perspective, challenges, and future research directions," Artif. Intell. Rev., vol. 55, pp. 453–563, 2022, doi: 10.1007/s10462-021-10037-y.

[4] G. Luo, Z. Chen, and B. O. Mohammed, "A systematic literature review of intrusion detection systems in the cloud-based IoT environments," Concurr. Comput. Pract. Exp., vol. 34, no. 15, Art. no. e6822, 2022, doi: 10.1002/cpe.6822.

[5] A. Alharthi, M. Alaryani, and S. Kaddoura, "A comparative study of machine learning and deep learning models in binary and multiclass classification for intrusion detection systems," Array, vol. 26, Art. no. 100406, 2025, doi: 10.1016/j.array.2025.100406.

[6] B. E. Hommel, F. M. Wollang, V. Kotova, H. Zacher, and S. C. Schmukle, "Transformer-based deep neural language modeling for construct-specific automatic item generation," Psychometrika, vol. 87, no. 2, pp. 749–772, 2022, doi: 10.1007/s11336-021-09823-9.

[7] L. Nie, Z. Li, J. Liu, and Y. Zhang, "Transformer-based network intrusion detection: A survey," ACM Comput. Surv., vol. 56, no. 2, pp. 1–38, 2024, doi: 10.1145/3606839.

[8] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.005.

[9] V. Buhrmester, D. Münch, and K. A. [last name assumed], "Analysis of explainers of black box models in a responsible machine learning framework," ACM SIGKDD Explor. Newsl., vol. 23, no. 1, pp. 55–65, 2021, doi: 10.1145/3468501.3468510.

[10] A. Das and M. J. Nene, "Explainable AI for intrusion detection systems: A survey," ACM Comput. Surv., vol. 55, no. 5, pp. 1–36, 2022, doi: 10.1145/3544547.

[11] R. Chinnasamy, M. Subramanian, E. Veerappampalayam, and J. Cho, "Deep learning-driven methods for network-based intrusion detection systems: A systematic review," ICT Express, vol. 11, pp. 181–215, 2025, doi: 10.1016/j.icte.2024.03.003.

[12] V. Hnamte and J. J. Hussain, "Dependable intrusion detection system using deep convolutional neural network: A novel framework and performance evaluation approach," Telem. Inform. Rep., vol. 11, Art. no. 100077, 2023, doi: 10.1016/j.teler.2023.100077.

[13] Y. Zhang, L. Wu, and S. Wang, "Multi-task learning for enhanced security analytics: A review and taxonomy," IEEE Trans. Dependable Secur. Comput., vol. 19, no. 5, pp. 2890–2908, Sep.–Oct. 2022, doi: 10.1109/TDSC.2021.3079979.

[14] E. Gyamfi and A. Jurcut, "Intrusion detection in Internet of Things systems: A review on design approaches leveraging multi-access edge computing, machine learning, and datasets," Sensors, vol. 22, no. 10, Art. no. 3744, 2022, doi: 10.3390/s22103744.

[15] Y. Wang, Z. Han, J. Li, and X. He, "BS-GAT behavior similarity based graph attention network for network intrusion detection," 2023, arXiv:2304.07226. [Online]. Available: https://doi.org/10.48550/arXiv.2304.07226

[16] H. Kheddar, "Transformers and large language models for efficient intrusion detection systems: A comprehensive survey," preprint, arXiv:2408.07583, 2024. [Online]. Available: https://arxiv.org/abs/2408.07583v1

[17] F. Ullah, A. Turab, S. Ullah, D. Cacciagrano, and Y. Zhao, "Enhanced network intrusion detection system for Internet of Things security using multimodal big data representation with transfer learning and game theory," Sensors, vol. 24, no. 13, Art. no. 4152, 2024, doi: 10.3390/s24134152.

[18] J. Mao, X. Yang, B. Hu, Y. Lu, and G. Yin, "Intrusion detection system based on multi-level feature extraction and inductive network," Electronics, vol. 14, no. 1, Art. no. 189, 2025, doi: 10.3390/electronics14010189.

[19] O. Arreche, T. Guntur, and M. Abdallah, "XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," Appl. Sci., vol. 14, no. 10, Art. no. 4170, 2024, doi: 10.3390/app14104170.

[20] R. Ahmed, R. A. Osman, and M. Amer, "Navigating urban congestion: A comprehensive strategy based on an efficient smart IoT wireless communication for PV powered smart traffic management system," PLOS ONE, vol. 19, no. 10, Art. no. e0310002, 2024, doi: 10.1371/journal.pone.0310002.

[21] K. S. Adewole, A. Jacobsson, and P. Davidsson, "Intrusion detection framework for Internet of Things with rule induction for model explanation," Sensors, vol. 25, no. 6, p. 1845, 2025.

[22] N. Kaur and L. Gupta, "Securing the 6G–IoT environment: A framework for enhancing transparency in artificial intelligence decision-making through explainable artificial intelligence," Sensors, vol. 25, no. 3, Art. no. 854, 2025, doi: 10.3390/s25030854.