XG-RAID: Explainable Graph Neural Network for Risk-Aware Intrusion Detection in IoT Ecosystems

Nasser ALJABRI[a], Mustafa AL SAMARA[a], Nabil Litayem[a], Abdelhak Belhi[a], Okba BEN ATIA[b], Ismail BENNIS[c]
[a]Cyber Defense Department, Joaan Bin Jassim Academy for Defence Studies, Qatar
[a]University of Technology of Belfort-Montbéliard, [c]IRIMAS, University of Haute-Alsace, France
{nh.aljabri, malsamara, abdelhak.belhi, nlitayem}@jbj.edu.qa, okba.ben-atia@utbm.fr, ismail.bennis@uha.fr

***Abstract*—The rapid expansion of the Internet of Things (IoT) has significantly increased the attack surface of modern networks, exposing heterogeneous devices to diverse cyber threats. While Graph Neural Networks (GNNs) can model complex traffic patterns, existing approaches often lack interpretability and operational risk awareness. This paper presents XG-RAID, an explainable graph-based intrusion detection framework that integrates relational learning with interpretability and risk-oriented analysis. Network flows from the CICIoT2023 dataset are represented as nodes in a similarity-based graph constructed using feature-space proximity. A GraphSAGE model performs multi-class classification across 23 attack categories. To enhance transparency, the framework incorporates SHAP, LIME, and Grad-CAM to provide feature-level and structural explanations. A risk scoring mechanism is introduced by combining prediction confidence with explanation consistency to prioritize alerts. Experimental results show that XG-RAID outperforms conventional machine learning models and a vanilla GraphSAGE baseline, while providing actionable insights for security analysts. Although the approach relies on similarity-based graph construction rather than explicit network topology, it offers a scalable and interpretable solution for IoT intrusion detection.**

***Index Terms*—Internet of Things (IoT), Intrusion Detection System (IDS), Graph Neural Networks (GNN), Explainable Artificial Intelligence (XAI), Network Security.**

I. Introduction

The Internet of Things (IoT) is transforming modern cyber-physical environments by interconnecting billions of heterogeneous devices across domains such as smart homes, healthcare, transportation, and industrial systems. While this connectivity enables automation and real-time monitoring, it also significantly expands the attack surface, exposing resource-constrained devices to threats such as Distributed Denial of Service (DDoS), botnets, spoofing, and reconnaissance attacks [1], [2].

Intrusion Detection Systems (IDS) play a central role in identifying malicious activities in IoT networks. However, traditional signature-based approaches struggle to detect novel or evolving attacks [3]. Classical Machine Learning (ML) methods improve generalisation but remain limited in highly dynamic IoT environments due to class imbalance, evolving traffic behaviour, and the lack of contextual modelling across related traffic samples [4]. As a result, these approaches often exhibit reduced robustness, particularly for low-frequency attack classes.

Graph Neural Networks (GNNs) have emerged as a promising approach for modeling dependencies in networked data by representing traffic samples as nodes in a graph structure [4], [5], enabling improved detection of distributed and multi-stage attacks. However, most GNN-based IDS operate as black-box models, limiting interpretability and hindering practical deployment. Conversely, Explainable Artificial Intelligence (XAI) techniques such as SHAP, LIME, and gradient-based methods improve model transparency [6], [7], but are typically applied to tabular or DL models and rarely integrated with graph-based intrusion detection. Moreover, existing works lack risk-oriented mechanisms to prioritize alerts based on both model confidence and interpretability.interpretability.

To address these limitations, we propose **XG-RAID**, an explainable and risk-aware graph-based intrusion detection framework for IoT environments. Network flows from the CICIoT2023 dataset are represented as nodes in a similarity-based graph constructed in feature space, enabling the model to capture statistical relationships between traffic patterns. A GraphSAGE architecture is used for multi-class classification across 23 traffic categories. To enhance transparency, the framework integrates SHAP, LIME, and Grad-CAM to provide complementary feature-level and structural explanations. In addition, a risk scoring mechanism is introduced by combining prediction confidence with explanation consistency to support alert prioritization. The main contributions of this paper are as follows:

- **Explainable graph-based IDS:** We propose a GNN-based intrusion detection framework that models IoT traffic using a similarity-driven graph representation and employs GraphSAGE for scalable multi-class classification.
- **Multi-level interpretability integration:** We integrate complementary XAI techniques (SHAP, LIME, and Grad-CAM) to provide both feature-level and structural explanations, improving transparency and analyst understanding.
- **Risk-oriented alert prioritization:** We introduce a practical risk scoring mechanism that combines prediction confidence with explanation consistency to support more informed security decision-making.

ISBN 978-3-903176-82-9 © 2026 IFIP

- **Evaluation on large-scale IoT data:** We validate the proposed framework on the CICIoT2023 dataset, demonstrating improved performance over conventional ML models and a vanilla GNN baseline while highlighting challenges related to class imbalance.

Unlike prior works that apply graph learning or explainability in isolation, XG-RAID provides a unified integration of relational modeling, interpretability, and risk-aware analysis within a single pipeline. While individual components exist in prior work, their integration into a unified risk-aware and explainable graph-based IDS remains limited. The rest of this paper is organised as follows. Section II reviews related work. Section III presents the proposed framework. Section IV describes the experimental setup. Section V discusses the results. Section VI concludes the paper.

II. Related Works

IDS for IoT environments have evolved significantly with the increasing scale and complexity of networked devices. Early approaches relied ML and DL models operating on flow-level features. While these methods improved detection performance, they often lacked contextual awareness and interpretability. Explainable ML-based IDS have been proposed to improve transparency.

In [6], traditional ML models were combined with LIME and SHAP to provide local and global explanations, enabling identification of influential features. Similarly, [3] introduced a DL-based IDS enhanced with RuleFit and SHAP to produce interpretable rule-based outputs. In [8], a comprehensive framework combined multiple ML models with several XAI techniques for feature importance analysis. Although these approaches improve interpretability, they treat network flows independently and fail to capture relationships between traffic samples, limiting their effectiveness in distributed IoT attack scenarios.

To address this limitation, recent work has explored GNNs for IoT intrusion detection. In [9], a GraphSAGE-based framework demonstrated improved classification performance by leveraging neighbourhood aggregation. Trust-aware extensions such as [5] further incorporated contextual trust information into message passing. Additionally, [2] proposed GNN architectures enhanced with temporal components and attention mechanisms, achieving strong detection accuracy in IoT environments. Despite these advances, most GNN-based approaches remain difficult to interpret and provide limited insight into model decisions, which restricts their usability in operational settings.

Existing approaches exhibit a trade-off between detection performance and interpretability: explainable ML methods provide transparency but lack relational modeling, while GNN-based models capture structural dependencies yet often operate as black-box systems. Moreover, few works incorporate risk-oriented analysis to prioritize alerts based on both prediction confidence and interpretability, particularly in large-scale and imbalanced IoT datasets such as CICIoT2023 [10]. To address these limitations, we propose XG-RAID, a unified framework that integrates graph-based learning with multi-level explainability and risk-aware analysis. By combining GraphSAGE with SHAP, LIME, and Grad-CAM, the approach enhances detection performance while providing interpretable and actionable outputs for IoT security environments.

III. The XG-RAID framework architecture

The proposed XG-RAID framework provides an explainable and risk-aware intrusion detection model for IoT ecosystems. As illustrated in Fig. 1, the framework consists of four stages: (i) data preprocessing, (ii) graph construction, (iii) GNN-based learning, and (iv) explainability and risk assessment.

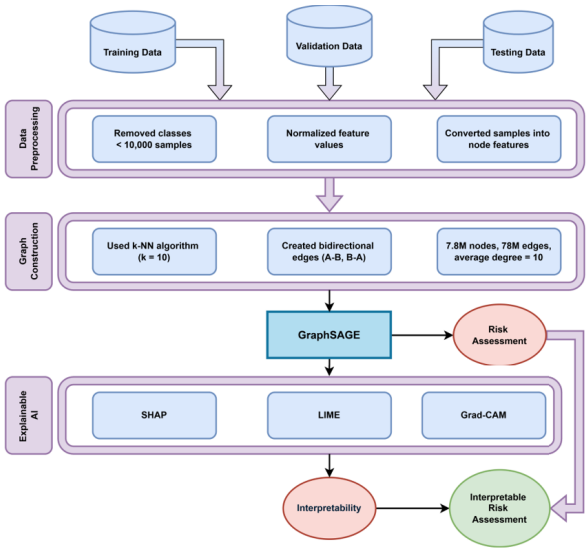

Fig. 1: Overview of the XG-RAID framework architecture.

A. Data Preprocessing

We evaluate XG-RAID on the CICIoT2023 dataset [10], which includes 23 traffic categories covering benign behaviour and multiple attack types such as DDoS, DoS, reconnaissance, spoofing, and botnet activity. Each network flow is represented by 86 numerical features describing statistical, packet-level, and temporal characteristics. To ensure stable training and graph connectivity, classes with fewer than 10,000 samples were excluded to ensure stable graph connectivity and avoid fragmented neighbourhood structures; this represents less than X% of the dataset and does not affect dominant attack patterns.

While this improves graph consistency, we acknowledge that rare attack detection remains critical and will be addressed in future work using techniques such as class-weighted learning and data augmentation. All features are normalized using min–max scaling, and the dataset is split into 70% training, 15% validation, and 15% testing sets. A traffic flow is defined as a bidirectional sequence of packets sharing the same 5-tuple (source IP, destination IP, source port, destination port, protocol) aggregated over a time window. Baseline models were configured using commonly adopted hyperparameters from prior IoT IDS literature.

B. Similarity-Based Graph Construction

Instead of relying on explicit network topology (e.g., IP-level communication graphs), XG-RAID constructs a *similarity graph* where nodes represent traffic flows and edges connect statistically similar samples in feature space. This design enables scalable graph construction for large datasets and facilitates the learning of shared behavioural patterns across flows. Edges are generated using a k-Nearest Neighbours (kNN) strategy:

- **Edge Formation:** Each node is connected to its $k = 10$ nearest neighbours based on Euclidean distance in the normalized feature space. The choice of $k = 10$ balances graph connectivity and computational cost, as validated in prior GNN-based IDS studies.
- **Symmetric Connectivity:** Edges are treated as bidirectional to form an undirected graph.

The resulting graph contains approximately 7.8 million nodes and 78 million edges. It captures statistical proximity between flows rather than explicit device interactions. This representation is particularly suitable for large-scale IoT datasets where direct communication metadata may be incomplete or unavailable. The proposed similarity-based graph differs from topology-aware graphs that explicitly model device communication. While topology-based approaches better capture physical interactions, they require packet-level metadata that is often unavailable in large-scale datasets such as CICIoT2023. In contrast, the similarity graph enables scalable construction while capturing behavioural proximity between flows, making it effective for detecting statistically similar attack patterns. However, it may not fully represent causal communication relationships, which will be explored in future hybrid graph designs.

C. Graph-Based Learning with GraphSAGE

GraphSAGE is an inductive GNN that learns node embeddings by aggregating features from sampled neighbourhoods, enabling scalability to unseen data. We adopt GraphSAGE [11] due to its scalability and inductive learning capability. For each node v, the embedding at layer k is computed as:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \text{AGG}\left(h_v^{(k-1)}, \{h_u^{(k-1)}, u \in \mathcal{N}(v)\}\right)\right) \quad (1)$$

where $\mathcal{N}(v)$ denotes the sampled neighbourhood, AGG is a mean aggregator, and σ is a non-linear activation function. Neighbourhood sampling ensures computational efficiency by limiting the number of neighbours processed per node. The final node embeddings are passed to a softmax classifier to predict one of the 23 traffic classes.

D. Explainability Integration

To provide interpretability, we integrate three complementary XAI techniques:

- **SHAP:** Applied on node feature vectors and model outputs to estimate global and local feature contributions. SHAP is computed in a model-agnostic manner using perturbations of input features while preserving the learned GNN predictions.
- **LIME:** Generates local surrogate models around each node by perturbing input features and approximating the GNN decision boundary, providing instance-level interpretability.
- **Grad-CAM for GNNs:** Adapted to graph structures by computing gradients of class scores with respect to node embeddings, highlighting influential nodes and neighbourhood regions contributing to predictions.

SHAP provides global feature attribution, LIME offers local decision approximation, and Grad-CAM highlights structural importance in graph neighborhoods. Their combination enables complementary multi-level interpretability.

E. Risk-Aware Scoring Mechanism

To support operational decision-making, we introduce a risk score that combines prediction confidence with explanation consistency:

$$R_i = \alpha \cdot P_i + (1 - \alpha) \cdot E_i \quad (2)$$

where P_i is the predicted probability of the assigned class for sample i, E_i represents explanation strength (e.g., normalized SHAP magnitude or explanation stability), and $\alpha \in [0, 1]$ is a weighting factor. This formulation allows the system to prioritize alerts that are both highly confident and strongly supported by explanations. Samples with conflicting or weak explanations are assigned lower risk scores and can be flagged for further inspection. In this work, E_i is computed as the normalized agreement between SHAP importance magnitude and LIME local fidelity, combined with Grad-CAM activation intensity:

$$E_i = \text{Norm}\left(\lambda_1 S_i + \lambda_2 L_i + \lambda_3 G_i\right) \quad (3)$$

The proposed risk score is not intended to replace classification decisions but to complement them by incorporating interpretability. High confidence predictions with weak or inconsistent explanations are down-weighted, reducing the risk of overconfident misclassifications. This is particularly relevant in imbalanced IoT scenarios, where rare attacks may produce uncertain predictions.

F. Scalability and Complexity Considerations

Given the large-scale nature of the dataset (7.8M nodes), scalability is a key consideration. GraphSAGE enables efficient mini-batch training through neighbourhood sampling, reducing computational complexity from $O(|E|)$ to approximately $O(B \cdot k^L)$, where B is the batch size, k is the number of sampled neighbours, and L is the number of layers. Memory usage is controlled by limiting neighbourhood size and using feature compression. The similarity-based graph construction can be parallelized using Approximate Nearest Neighbour (ANN) methods, making the approach feasible for large-scale IoT datasets. Overall, XG-RAID provides a scalable and interpretable framework that balances detection performance with practical deployment constraints. The above complexity refers to training, while inference scales linearly with the number of evaluated nodes.

IV. Evaluation Setup

This section presents the experimental setup used to evaluate the XG-RAID framework, including baseline models, training configuration, and evaluation metrics.

A. Baseline Models and Training Configuration

To assess the effectiveness of XG-RAID, we compare it against widely used baseline models in IoT intrusion detection:

- **Random Forest (RF):** a robust baseline for anomaly detection on tabular data.
- **XGBoost:** a strong gradient boosting model with high performance on structured datasets.
- **Deep Neural Network (DNN):** a fully connected architecture commonly used in IoT IDS [3].
- **GraphSAGE (vanilla):** a graph-based baseline without explainability or risk-aware components.

These baselines represent both tabular and graph-based learning paradigms, enabling a comprehensive comparison with the proposed approach. The XG-RAID model is implemented using a two-layer GraphSAGE architecture with a mean aggregator and a hidden dimension of 128. Training is performed with a batch size of 4,096 nodes, a learning rate of 1×10^{-3}, the Adam optimizer, and a dropout rate of 0.2 over 50 epochs. Experiments are conducted on a GPU-enabled workstation to support scalable mini-batch training on large graph structures. The reported training configuration reflects a trade-off between scalability and performance, ensuring that the proposed framework remains feasible for large-scale IoT deployments. Quantitative scores (completeness, fidelity, and relevance) are used as proxy measures of explanation quality.

B. Evaluation Metrics

To evaluate classification performance, we employ standard multi-class metrics, along with additional measures suited for imbalanced intrusion detection scenarios. Let TP, FP, TN, and FN denote True Positives, False Positives, True Negatives, and False Negatives, respectively.

- **Accuracy (ACC):** Represents the proportion of correctly classified samples over the entire dataset.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- **Precision:** Measures how many predicted positive samples are truly positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- **Recall (True Positive Rate, TPR):** Quantifies the proportion of actual positives correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- **F1-score:** The harmonic mean of precision and recall.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

- **False Positive Rate (FPR):** Represents the proportion of benign samples incorrectly classified as attack.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{8}$$

- **False Negative Rate (FNR):** Measures the proportion of attack samples that are incorrectly classified as benign.

$$\text{FNR} = \frac{FN}{TP + FN} \tag{9}$$

- **AUC-ROC:** To measure class separability, we compute the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). For a given class c, AUC is defined as:

$$\text{AUC}_c = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t)) \tag{10}$$

  where t is the decision threshold, and The macro-AUC is the unweighted average of AUC_c across all classes.

For multi-class evaluation, precision, recall, F1-score, FPR, and FNR are computed per class and then macro-averaged. The inclusion of FPR and FNR is particularly important in IoT intrusion detection, where false alarms and missed attacks have significant operational impact, especially under class imbalance. A confusion matrix M is also used to analyse class-level misclassifications, where each entry $M_{i,j}$ represents the number of samples of class i predicted as class j.

C. Interpretability and Risk Evaluation

In addition to classification performance, we assess interpretability and risk-awareness:

- **SHAP completeness:** evaluates how well feature attributions explain the model output.
- **LIME local fidelity:** measures the accuracy of local surrogate models in approximating GNN predictions.
- **Grad-CAM structural relevance:** assesses whether highlighted graph regions correspond to meaningful traffic patterns.

Finally, the proposed risk scoring mechanism is evaluated qualitatively based on its ability to produce consistent and interpretable prioritization of alerts, particularly in ambiguous or low-frequency attack scenarios.

V. Results and Discussion

This section evaluates the performance of XG-RAID on the CICIoT2023 dataset, focusing on classification accuracy, robustness under class imbalance, explainability, and risk-aware decision support.

Table I presents the overall classification results. XG-RAID achieves the highest performance across all metrics, with an accuracy of 96.4% and an F1-score of 95.3%, outperforming both classical ML models and the vanilla GraphSAGE baseline. The improvement over vanilla GraphSAGE (+1.7% accuracy) indicates that the proposed framework benefits from the combined effect of similarity-based graph construction and enhanced decision-level analysis. Compared to tabular models

TABLE I: Overall Multi-Class Classification Performance

Model	Accuracy	Precision	Recall	F1-score
RF	89.4%	87.2%	85.9%	86.0%
XGBoost	92.1%	90.8%	89.7%	90.1%
DNN	90.6%	89.2%	88.1%	88.6%
GraphSAGE (vanilla)	94.7%	93.5%	92.8%	93.1%
XG-RAID	**96.4%**	**95.6%**	**95.1%**	**95.3%**

such as XGBoost and RF, the results highlight the advantage of incorporating relational context between traffic samples.

Table II reports macro-averaged performance across major attack families. As expected, high-frequency attacks such as DDoS/DoS achieve near-perfect detection, while performance decreases for less frequent categories such as Botnet behaviour. From an error perspective, high-frequency classes exhibit low FPR and FNR, indicating stable detection with minimal false alarms. In contrast, rare attack classes show increased FNR, reflecting missed detections due to limited representation. This behaviour is consistent with prior IoT intrusion detection studies and highlights the importance of complementary risk-aware analysis.

TABLE II: Per-Attack-Type Classification Performance (Macro-Averaged)

Attack Family	Precision	Recall	F1-score
DDoS / DoS	97.8%	96.9%	97.1%
Reconnaissance	95.6%	94.8%	95.1%
Botnet	92.3%	90.7%	91.3%
Overall Macro	**95.2%**	**94.1%**	**94.5%**

Although the full confusion matrix is omitted due to space constraints, class-level analysis reveals two consistent trends:

- **High separability of volumetric attacks:** DDoS-related classes are clearly distinguished due to strong statistical signatures.
- **Overlap among low-frequency classes:** Certain Botnet and reconnaissance variants exhibit partial misclassification due to similar behavioural patterns in feature space [6], [9].

These observations indicate that errors mainly concentrated in structurally similar and underrepresented classes.

As shown in Fig. 2, the proposed GraphSAGE-based model achieves high AUC values for dominant attack categories such as DDoS and DoS, indicating strong class separability. In contrast, lower AUC scores observed for rare attack classes reflect the impact of class imbalance in the dataset.

Fig. 3 presents the Precision–Recall (PR) curves for the GraphSAGE model across the 23 traffic classes. PR-AUC provides a more informative evaluation under class imbalance by emphasizing the trade-off between precision and recall. As shown in Fig. 3, the proposed model maintains high PR-AUC values for frequent attack categories such as DDoS and DoS, indicating reliable detection with low false-alarm rates. In contrast, reduced PR-AUC scores for rare attack classes highlight the inherent difficulty of identifying low-frequency behaviours in highly imbalanced IoT datasets.

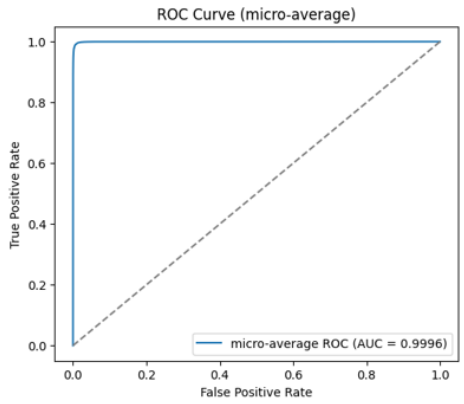

Fig. 2: ROC-AUC curves for the GraphSAGE model across the 23 traffic classes.

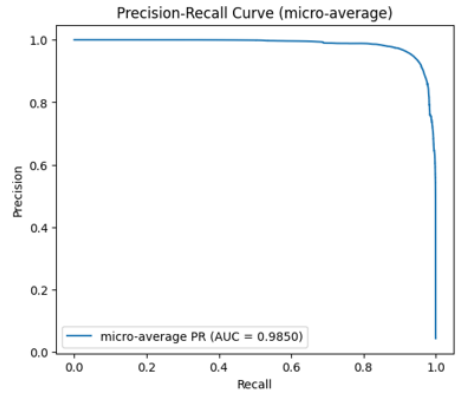

Fig. 3: PR-AUC curves for GraphSAGE across all traffic classes under class imbalance.

These results further motivate the risk-aware analysis adopted in XG-RAID, where attack severity and class frequency are jointly considered beyond aggregate accuracy. As stated previously, XG-RAID incorporates three complementary XAI techniques. For example, a DDoS sample shows high SHAP importance for packet rate and strong Grad-CAM activation in dense neighborhoods, confirming consistent multi-level explanations. Below, we summarise the key explainability results. The **SHAP** identifies key features such as flow inter-arrival time, packet length variance, and SYN/ACK ratio as dominant contributors. The high completeness score (0.87) indicates that feature attributions capture most of the model decision process. The **LIME** Achieves a high local fidelity (0.92), confirming that local surrogate models closely approximate GNN predictions at the instance level. The **Grad-CAM** Reveals distinct structural patterns across attack types, including dense neighbourhood activations for DDoS traffic and sparse structures for Botnet behaviour.

As illustrated in Fig. 4, Grad-CAM reveals distinct node-level activation patterns across traffic classes. Volumetric attacks such as DDoS and DoS (e.g., Class 18 – DDoS-UDP) exhibit strong contribution values concentrated in dense node

neighborhoods, reflecting coordinated high-volume traffic behaviour. In contrast, benign traffic (Class 1) shows low and diffuse activation values, indicating the absence of a dominant structural signature. Reconnaissance activities present more distributed, low-density activation patterns, while Botnet behaviours are characterized by sparse and isolated node clusters. These observations confirm that Grad-CAM effectively captures class-specific structural characteristics within the graph.

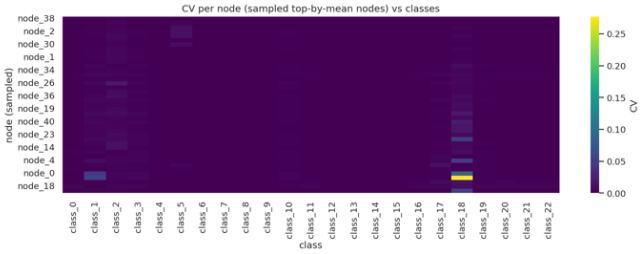

Fig. 4: Grad-CAM Node-Wise Contribution Value Heatmap

Although a full ablation study is beyond the scope of this work, the observed improvement over vanilla GraphSAGE suggests that graph construction and decision-level enhancements contribute more significantly to performance than individual components alone. XG-RAID generates a combined risk score by merging prediction confidence with explanation strength. Key findings include:

- **High-risk alerts (confidence $>$ 0.90):** Mostly associated with volumetric DDoS traffic, showing clear feature-level and graph-level justification.
- **Medium-risk alerts:** Often linked to reconnaissance variants where SHAP and LIME present competing feature influences.
- **Low-risk alerts:** Typically correspond to benign traffic or overlapping low-frequency attack classes.

This categorization assists analysts in prioritizing incidents and reduces false alarms by incorporating interpretability into the decision-making process.

The experiments demonstrate that XG-RAID achieves superior performance compared to classical ML models and standard GNN architectures, while also handling highly imbalanced datasets more effectively than existing baselines. In addition to improved accuracy, XG-RAID provides actionable and interpretable explanations by revealing both feature-level influences and graph-structural patterns that drive model decisions. The framework further generates meaningful risk scores that reflect prediction uncertainty and explanation consistency, enabling analysts to prioritise alerts with greater confidence. These strengths collectively illustrate the practicality and robustness of XG-RAID for real-world IoT intrusion detection, particularly in operational environments requiring transparency, trust, and accountability.

Despite strong overall performance, the proposed approach exhibits reduced sensitivity for low-frequency attack classes, as reflected by higher FNR values. Additionally, the similarity-based graph does not explicitly model communication topology, which may limit detection of certain coordinated attacks. These limitations highlight directions for future improvements.

VI. Conclusion

This paper presented XG-RAID, an explainable and risk-aware intrusion detection framework for IoT environments. By modeling network traffic as a similarity-based graph and leveraging GraphSAGE, the approach captures contextual relationships between flows, improving multi-class attack detection. The integration of SHAP, LIME, and Grad-CAM enables multi-level interpretability, providing feature-level and structural insights to support analyst decision-making. In addition, a risk scoring mechanism combines prediction confidence with explanation consistency to prioritize alerts in operational settings. Experimental results on CICIoT2023 show that XG-RAID outperforms conventional ML models and a vanilla GNN baseline while delivering interpretable outputs. However, the reliance on similarity-based graphs and reduced performance on low-frequency attacks remain limitations. Future work will explore topology-aware graph modeling, improved handling of rare attacks, and lightweight GNN designs for real-time deployment.

References

[1] Alsbatin, L., Zawaideh, F., Alrifai, B.M. and Alawneh, T.A., 2025. Enhancing Internet of Things (IoT) Network Security: A Machine Learning-Driven Framework for Real-Time Intrusion Detection and Anomaly Classification. Mesopotamian Journal of CyberSecurity, 5(3), pp.1042-1056.

[2] Rivera, A. and Uribe, J., 2025, June. Graph based machine learning for anomaly detection in iot security. In ECCSUBMIT Conferences (Vol. 3, No. 2, pp. 40-48).

[3] Abou El Houda, Z., Brik, B. and Senouci, S.M., 2022. A novel IoT-based explainable deep learning framework for intrusion detection systems. IEEE Internet of Things Magazine, 5(2), pp.20-23.

[4] Altaf, T., Wang, X., Ni, W., Yu, G., Liu, R.P. and Braun, R., 2024. GNN-based network traffic analysis for the detection of sequential attacks in IoT. Electronics, 13(12), p.2274.

[5] Awan, K.A., Din, I.U., Almogren, A., Han, Z. and Guizani, M., 2025. TrustAware-GNN: Graph Neural Network-Based Trust Management for IoT Anomaly Detection. IEEE Internet of Things Journal.

[6] Çelik, A.F., Sağlam, B. and Demirci, S., 2023, October. Developing Explainable Intrusion Detection Systems for Internet of Things. In 2023 16th International Conference on Information Security and Cryptology (ISCTürkiye) (pp. 1-6). IEEE.

[7] Adhikari, D. and Thapaliya, S., 2024. Explainable AI for cyber security: interpretable models for malware analysis and network intrusion detection. NPRC Journal of Multidisciplinary Research, 1(9), pp.170-179.

[8] Gummadi, A.N., Napier, J.C. and Abdallah, M., 2024. XAI-IoT: an explainable AI framework for enhancing anomaly detection in IoT systems. IEEE Access, 12, pp.71024-71054.

[9] Bibi, I., Ozcelebi, T. and Meratnia, N., 2023, October. An IoT Attack Detection Framework Leveraging Graph Neural Networks. In International Conference on Intelligence of Things (pp. 225-236). Cham: Springer Nature Switzerland.

[10] Himadri. (2025). CICIoT2023: A Modern Dataset for Intelligent IoT Threat Detection [Data set]. MIT License. Kaggle. https://www.kaggle.com/datasets/himadri07/ciciot2023

[11] Hamilton, W., Ying, Z. and Leskovec, J., 2017. Inductive representation learning on large graphs. Advances in neural information processing systems, 30.

[12] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.