International Journal of Applied Mathematics

Volume 38 No. 4s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

INTRUSION DETECTION IN CYBERSECURITY: A STUDY ON EXPLAINABLE GRAPHIC REINFORCEMENT LEARNING

Arun Kumar B S[1*] , Rathnakar Achary[2]

[1]Research Scholar, Department of Computer Science and Engineering, Alliance School of Advanced Computing, Alliance University, Bangalore, Karnataka- 562106 India

*Corresponding Email Id: karunphd23@ced.alliance.edu.in

[2]Professor, Department of Computer Science and Engineering, Alliance School of Advanced Computing, Alliance University, Bangalore, Karnataka- 562106 India

Email Id: rathnakar.achary@alliance.edu.in

Abstract

Intrusion Detections Systems (IDS), which are consequently vital for safeguarding digital infrastructure, counter evolving cyber threats. Often, conventional IDS systems including signature-based and anomaly-based battle dynamic attack patterns and high false warning rates. Artificial intelligence (AI) driven solutions, especially reinforcement learning (RL) and graph-based models—have grown more popular in reaction to their capacity to adapt and identify sophisticated threats. As a result, the lack of transparency that is associated with AI-driven intrusion detection systems provides a significant challenge for decision-makers in the field of cybersecurity. Growing confidence and interpretability in AI-based intrusion detection have been greatly influenced by explainable artificial intelligence (XAI). Emphasizing their efficacy in modeling network traffic, enhancing detection accuracy, and guaranteeing decision transparency, this paper seeks to investigate the incorporation of explainability in graph-based reinforcement learning models for IDS. Using secondary data gathering from online databases covering the years 2018 to 2025, a qualitative research approach is employed. The study methodically surveys research on explainability methods in AI-driven IDS, graph-based intrusion detection, and reinforcement learning applications in cybersecurity. Though explainability systems increase interpretability with minimal accuracy loss, the results show that graph-based RL improves intrusion detection and network traffic analysis by utilizing structural links. Nevertheless, problems including adversarial assaults, computation costs, and the trade-off between openness and performance remain. The research shows that using explainable artificial intelligence in graph-based RL IDS can significantly increase detection capabilities and user confidence, hence promoting more efficient and responsible cybersecurity solutions, future studies should concentrate on increasing the scalability, durability, and real-time applicability of explainable graph-based RL models in the field of cyber security.

Keywords: *Explainable Artificial Intelligence; Intrusion Detection Systems (IDS); Artificial Intelligence (AI); Reinforcement Learning (RL); Cybersecurity; Explainable Graph*

Received: August 02, 2025

Introduction

As cyber-attacks have grown in complexity and frequency, cyber security has emerged as a major concern in the modern digital landscape. Often inadequate against advanced persistent threats (APTs), polymorphic malware and zero-day vulnerabilities [1, 2], conventional security measures like firewalls and antivirus software fall short. By way of network traffic monitoring, identification of hostile activities, and alerting of security staff of potential hazards, IDS are a vital defensive tool. Notable disadvantages of conventional IDS techniques, including signature-based and anomaly-based detection, exist [2, 3]. While anomaly-based Intrusion Detection Systems generate higher false positive rates because of their reliance on deviations from normal traffic patterns, Signature-based Intrusion Detection Systems rely on established attack patterns, hence rendering them worthless against new threats [3, 4]. Attacks' complexity drives the need for sophisticated detection techniques able to dynamically adapt to new threats while maintaining accuracy and efficiency [4]. Without rule sets, RL is a potent cybersecurity tool since it can learn and adjust to fresh attack patterns. Reinforcement learning-based intrusion detection systems can identify and mitigate threats by treating network security as a sequential decision-making problem [5, 6]. Graphical models of network traffic can clarify structural relationships between nodes and identify uncommon interactions faster than flat-feature models [7, 8]. Combining graph neural networks (GNNs) with RL increases attack detection by leveraging network data spatial and temporal correlations. Though they have promise, artificial intelligence-driven intrusion detection systems lack transparency. Many deep learning-based security systems run as "black boxes," which makes it challenging for cybersecurity experts to know, assess, and depend on their choices [9]. By providing human-interpretable insights into IDS decisions, XAI helps to close this gap by increasing confidence and supporting regulatory compliance. The figure 1 illustrates the trade-off in AI/XAI-based cybersecurity between three key aspects.

This review article aims to investigate the integration of explainable artificial intelligence into graph-based reinforcement learning models for intrusion detection, hence assessing their effectiveness in modelling network traffic, improving detection accuracy, and guaranteeing decision transparency. The paper is structured as follows: Beginning with a presentation of conventional Intrusion Detection Systems (IDS) and their related problems, the paper then thoroughly examines graph-based learning and reinforcement learning in the framework of cybersecurity. Then, it evaluates performance and trust elements and studies explainability techniques in AI-driven Intrusion Detection Systems. The report points out research gaps and offers next paths to enhance the usability of explainable graph-based reinforcement learning models in cybersecurity.

Received: August 02, 2025

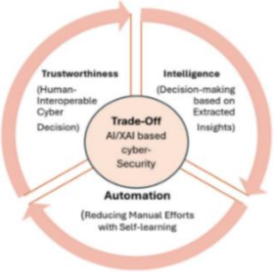
Trustworthiness
(Human-Interoperable Cyber Decision)
Intelligence
(Decision-making based on Extracted Insights)
Trade-Off
AI/XAI based cyber-Security
Automation
(Reducing Manual Efforts with Self-learning

Figure 1. Effective and explainable cybersecurity solutions – An overview [9]

Intrusion Detection Systems – An Overview

Cybersecurity depends on IDS, which provide tools to spot illicit access and dangerous actions. Signature-based IDS identifies known threats using known attack signatures; anomaly-based IDS detects abnormalities from normal behavior using statistical or AI-driven models; hybrid IDS combines both methods to increase detection accuracy and versatility [10]. Although anomaly-based IDS can identify new threats, it may generate more false positives; signature-based IDS is excellent at locating current threats but weak against zero-day attacks. Traditional Intrusion Detection Systems face many challenges, including scalability constraints, the changing nature of attack vectors, and an increasing amount of network traffic. As cyber threats grow in complexity, static rule-based models become ineffective; adaptive learning models that can develop with emerging threats are therefore required. AI and machine learning (ML) methods have been applied in IDS as a result, hence facilitating quick decision-making and automated threat identification. Though there have been developments, traditional machine learning-based intrusion detection systems have drawbacks like data imbalance, susceptibility to hostile attacks, and a lack of explainability. Many machine learning models operate as black-box systems, hence perplexing the reading of detection outcomes for security professionals. Moreover, to stay successful against evolving threats, static machine learning models require regular retraining [10, 11, 12]. These deficiencies draw attention to the significance of more strong, clear, and flexible Intrusion Detections Systems architectures—including reinforcement learning and graph-based approaches—which can continuously learn from new threats while maintaining interpretability and dependability in cybersecurity applications.

Graph-Based Approaches in Intrusion Detection

Graph-based approaches have emerged as a powerful paradigm in IDS, using the natural structural properties of network traffic to enhance threat identification. Cybersecurity depends on graph theory since it uses organized graphs to model complex network interactions, with nodes representing people, equipment, or IP addresses and edges indicating data flows or communication links. This strategy helps to catch contextual links and dependencies sometimes overlooked by more traditional approaches. Unlike traditional IDS, which examine network behaviour in isolation, graph-based methods assess the overall structure of network traffic, therefore allowing the detection of complex and dynamic attack patterns [6, 7]. Graph analytics allows security systems to more quickly identify anomalies, map attack propagation, and follow enemy movements inside a network than signature-based or purely statistical approaches.

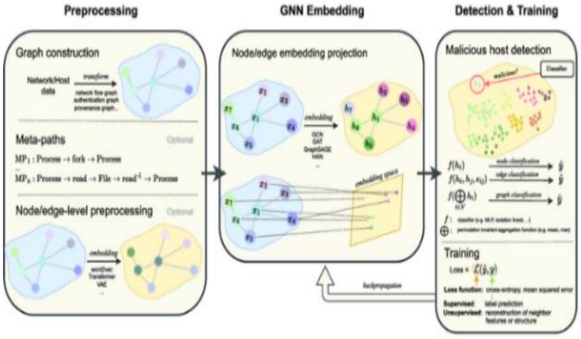

Figure 2. General architecture for intrusion detection with GNN-based methods[1]

Graph-based IDS relies fundamentally on the depiction of network traffic as graphs. This means turning raw network data into an ordered representation where nodes represent network entities—servers, users, applications—and edges indicate interactions like data transfers, authentication attempts, or session connections. This organized depiction facilitates enhanced understanding of traffic patterns, aiding in the differentiation of typical interactions from possible hazards. Moreover, dynamic graphs allow real-time monitoring and modification of developing attack methods by means of temporal evolution of edges and nodes. By detecting patterns from graph-structured data, GNNs enhance this capacity and enable the automatic detection of anomalies including botnet operations, data exfiltration, or APTs [7, 13]. These models include methods including "Graph Convolutional Networks (GCNs)", "Graph Attention Networks (GATs)'', and Graph Auto encoders to uncover latent links and draw possible detrimental actions from relational patterns. The figure 2 below illustrates the General architecture for intrusion detection with GNN-based methods.

The comparative advantages of graph-based intrusion detection over conventional methods include its capacity to capture non-linear dependencies, predict developing threats, and diminish false positives. Conventional machine learning-based intrusion detection systems depend on predetermined feature sets and frequently falter against adversarial attacks, but graph-based solutions utilize structural relationships and contextual awareness, rendering them more resilient. Moreover, GNNs facilitate semi-supervised learning, permitting detection models to function well despite the scarcity of labelled data. A significant benefit is the scalability of graph-based Intrusion Detection Systems, as network security graphs can effectively manage extensive network topologies [13, 14]. Nonetheless, despite their advantages, obstacles include processing cost, the necessity for real-time adaptation, and the intricacies of explainability persist as subjects of continuing investigation. Graph-based

[1] https://www.researchgate.net/figure/General-architecture-for-intrusion-detection-with-GNN-based-methods-In-a-first_fig1_370733865

methodologies signify a revolutionary advancement in intrusion detection, improving detection precision, flexibility, and robustness against contemporary cyber threats.

Reinforcement Learning for Intrusion Detection

Reinforcement Learning provides a robust foundation for intrusion detection by facilitating adaptive and autonomous security mechanisms against dynamic cyber threats. Reinforcement Learning fundamentally relies on Markov Decision Processes (MDP), wherein an agent engages with an environment to ascertain best actions via trial and error. The agent obtains rewards or punishments contingent upon its behaviours, thereby refining its policy over time to optimize long-term security results. The self-learning feature of reinforcement learning renders it especially appropriate for intrusion detection, as cyber threats continually evolve, necessitating dynamic and context-sensitive countermeasures [5, 6]. In contrast to conventional static IDS models, RL-based IDS may adjust to novel attack patterns, optimize resource distribution, and improve detection precision without depending exclusively on predetermined signatures or labelled datasets. The figure 3 illustrates the overview of reinforcement learning based IDS in detail.

By enabling real-time decision-making and anomaly detection in large networks, Deep RL, which combines RL with deep neural networks, improves cyber security. DRL-based intrusion detection can independently categorize threats, prioritize notifications, and execute pre-emptive responses by simulating attack-defence situations. These systems' foundation is the Q-learning based algorithms:

"$Q(s_t, a_t) \leftarrow \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1,a}) - Q(s_t, a_t)$" ----------------------------Eq. (1)

Where

$Q(s_t, a_t)$ – It is the Q-value for the state s_t and action a_t.

α - It is the learning rate

r_{t+1} – It is the reward received after taking action a_t.

γ - It is the discount factor, which determines the importance of future rewards

$\max_a Q(s_{t+1,a})$ -It is the estimated maximum future rewards for the next state s_{t+1}

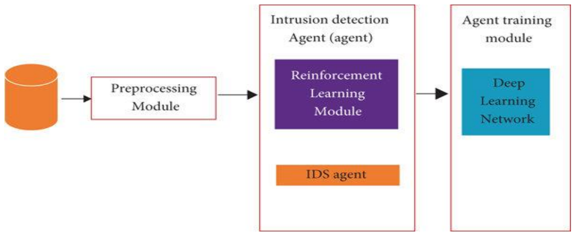

Figure 3. Reinforcement learning based IDS[2]

[2] https://www.researchgate.net/figure/Proposed-reinforcement-learning-based-IDS-for-IOT-data_fig1_359075718

By choosing actions depending on the existing network sate St (represented by the graph embedding), the RL agent is charged with dynamically identifying threats. To maximize long -term rewards, the agent discovers an optimal strategy, π^* (a|S_t). Which is

$\pi^*(a|S_t)$ to maximize the long -term rewards. Where:

"$\pi^*(a|S_t) = \arg_\pi \max E\,[R_t|S_t, \pi]$" ------------------------------------ Eq (2)

Furthermore, reinforcement learning-driven adaptive defensive mechanisms improve intrusion detection system effectiveness by dynamically modifying detection thresholds, revising security policies, and implementing countermeasures according to the intensity of an assault. Nonetheless, implementing reinforcement learning in cybersecurity entails problems like training complexity, processing demands, exploration-exploitation dilemmas, and adversary interference [15]. Despite these difficulties, RL-based IDS can improve real-time threat mitigation, network resilience, and security concerns in modern cybersecurity systems.

Explainability in Ai-Driven Intrusion Detection

AI-driven intrusion detection must be explainable to improve cybersecurity decision-making transparency, reliability, and interpretability. Cybersecurity specialists struggle to understand warning triggers as AI-driven intrusion detection systems (IDS) get more complex. Their decision-making procedures typically resemble black-box models. Insufficient interpretability can lead to false positives, missed threats, and regulatory compliance issues [16]. Explainability helps security analysts validate threat classes, investigate warnings, and develop detection models to improve network security by revealing model behaviour.

Two primary techniques for explainability are post-hoc and intrinsic. Post-hoc explainability techniques such as SHAP and LIME retrospectively assess model forecasts. While LIME approximates complex models with simpler, understandable models to clarify local decision-making, SHAP assigns relevance scores to input features highlighting their impact on categorization. Moreover, deep learning model attention mechanisms highlight important input areas affecting an IDS choice, thereby improving analyst interpretation [16, 17]. These approaches allow specialists in cybersecurity to verify, diagnose, and improve AI-driven Intrusion Detection Systems, hence ensuring that the responses are clear and accurate.

The increasing fascination in Explainable artificial intelligence (XAI) arises from the need to render machine learning models more open. XAI techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) provide ways to understand the forecasts of black-box models. XAI can enable security analysts to know why a particular action was taken in the framework of RL-based IDS, hence fostering confidence in the system. SHAP uses the Shapley value defined as:

"$\emptyset_i(f) = \sum_{s \subseteq N\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$" ---------------------------Eq (3)

Where,

$\emptyset_i(f)$ – It is the Shapley value for feature i,

S – It is a subset of the features excluding i,

N - It is the set of all features,

Techniques like SHAP are included in XAI to clarify the choices of the RL agent. For every choice, the algorithm calculates feature importance ratings indicating which elements—such network characteristics—most influenced the identification of harmful activity.

"$\emptyset_i(f) = \sum_{s \subseteq N\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$"-------------------------- Eq (4)

On the other hand, models especially in graph-based and RL-driven IDS include natural intrinsic explainability. Presenting a natural framework for intrusion detection, GNNs capture relational interdependence across network nodes and edges and provide understandable insights on connection strength and node relevance. RL models can also be designed with explained reward functions, boosting the transparency of their decision-making processes [17]. Still, often explainability means sacrificing precision and economy. While very accurate deep learning models lack openness, more understandable models could sacrifice performance. The evolution of AI-driven Intrusion Detections Systems that are both successful and responsible in cybersecurity defense depends on finding balance among detection accuracy", computation efficiency, and interpretability.

Integration of Graph-Based Learning with Reinforcement Learning for ids:

Combining graph-based learning with RL for IDS creates a novel approach for adaptive threat mitigation and real-time attack identification. With nodes signifying devices, edges expressing connections, and relational dependencies showing behavioural patterns, GNNs shine at describing complex network topologies. The result is a more dynamic and context-sensitive intrusion detection system competent of proactively reducing incursions when combined with reinforcement learning, which continuously learns and adapts to new threats. Using GNNs to produce sophisticated graph representations of network traffic and including these representations into RL-based decision models allows IDS to dynamically change detection rules, forecast hostile behaviours, and improve countermeasures in real time. GNNs' message-passing mechanism modifies node characteristics depending on their neighbours:

"$h_v^{(k)} = \sigma\left(W^{(k)}. \sum_{u \in N(v)} h_u^{k-1} + b^{(k)}\right)$" ---------------------------------- Eq (5)

Where,

$h_v^{(k)}$ – It is the feature vector of node v adter k^{th} interation (or layer),

$W^{(k)}$ – It is the weight matrix for layer k,

N(v) – It represents the neigbbours of node v,

Received: August 02, 2025

Volume 38 No. 4s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

h_u^{k-1} – It is the feature vector of neighbor u from the previous layer,

σ -It is the activation funciton (e.g., ReLU)

$b^{(k)}$ - It is the bias term for layer k,

As defined by, the GNN updates node features depending on nearby node information by means of message passing between nodes.

"$h_v^{(k)} = \sigma\left(W^{(k)}.\sum_{u\in N(v)} h_u^{k-1} + b^{(k)}\right)$" ------------------------------- Eq (6)

Where,

$h_v^{(k)}$ – It is the updated feature vector of node v layer k

$W^{(k)}$ and $b^{(k)}$ - They are the trainable weight and bias matrices

N(v It represents the neighbors of node v

Graph-based RL enables intelligent network traffic monitoring and adaptive threat mitigation by means of the conceptualizing of intrusion detection as a sequential decision-making problem. Using relational data, the IDS agent monitors the network graph, finds anomalies, and computes suitable countermeasures by means of trial and error learning. Reinforcement learning methods, especially deep Q-networks (DQNs) and policy-gradient approaches, may enhance defence measures by means of network state transition analysis. Ensuring that security rules can change in reaction to new threats helps to achieve this. Graph embedding's use in reinforcement learning improves the generalization ability of the model, which thus allows intrusion detection systems to identify new attacks more efficiently than is feasible with conventional signature-based approaches [18]. Though it offers certain benefits, the combination of GNNs with RL raises several issues. The notable computational complexity is a major issue since both graph-based learning and reinforcement learning need a great deal of resources for training and inference.

Moreover, keeping large networks made up of millions of nodes and edges causes scale problems that render real-time analysis a resource-consuming task. Reward engineering—the design of an acceptable reward function for intrusion detection—is a major difficulty. This function has to balance system performance, false positive rates, and detection accuracy [18]. Finally, adversarial assaults on graph topologies and reinforcement learning rules create a security concern and call for strong defenses to prevent model exploitation. Facing these obstacles is essential to completely exploit the possibilities of graph-based reinforcement learning for the creation of solutions for next-generation intrusion detection systems.

METHODOLOGY AND COMPARATIVE ANALYSIS OF VARIOUS STUDIES

This table now offers exact accuracy rates as cited in the study, so enabling a more obvious comparison of the effectiveness of several approaches.

Table 1. Comparative Analysis

References	Methodology	Findings	Limitations
Thang & Pashchenko (2019)	Multistage ML-based IDS for WiFi networks	Achieved **98.9% detection accuracy** for WiFi intrusions	Limited to WiFi networks; lacks adaptability to emerging attack patterns
Palmer, Rogers & Mcfly (2020)	Graph-based study of industrial control system (ICS) network traffic	Detected **88.2%** of anomalous behaviours in ICS networks	Lacks real-time detection capabilities
Abou Daya et al., (2020)	ML-based graph-based bot detection (BotChase)	Identified botnet (Botchase) activities with **99% accuracy**	Large-scale networks' high computational cost
Neupane et al., (2022)	Survey on explainable IDS (X-IDS)	Provided a comparative analysis of existing XAI methods; noted that most models maintain explainability at the cost of **5-10% accuracy drop**	No empirical validation of proposed methodologies
Baahmed et al. (2023)	GNN for intrusion detection method and the explanation	GNN-based IDS achieved **99.54% accuracy** with improved interpretability	Model interpretability and explainability trade-offs remain a challenge
Lo (2023)	Graph representation learning for cyberattack detection	Enhanced forensic analysis and attack attribution; increased detection rate to **99.54%**	Requires large datasets for effective learning
Kaya et al. (2024)	X-CBA: Explainability-aided CatBoost model for IDS	Achieved **99.47% detection accuracy**, improving interpretability in decision-making	Explainability performance in complex cyberattacks not fully assessed
Adhikari & Thapaliya (2024)	Explainable AI (XAI) models for malware and	XAI-based models improved interpretability	Focuses on theoretical concepts rather than real-world deployment

	intrusion detection	while maintaining **80% accuracy**	
Farrukh et al. (2024)	Xg-NID: Heterogeneous graph neural network with LLM for IDS	Demonstrated **97.2% detection accuracy** by integrating multimodal data	High computational complexity for large-scale deployment
Shokouhinejad et al. (2025)	Graph learning and XAI for malware detection	Combined graph learning and explainability, achieving around **94% classification accuracy.** Graph reduction and embedding techniques have tackled issues with scalability and efficiency, whilst explainability has connected high detection accuracy with actionable insights.	Trade-off between detection performance and explainability
Kalutharage et al. (2025)	The combination of neurosymbolic learning and domain knowledge-driven explainable artificial intelligence for Internet of Things attack detection and response	Achieved **97.1% detection accuracy**, improving interpretability and response efficiency in IoT networks	Increased computational complexity and dependency on high-quality domain knowledge for effective reasoning
Ahanger et al. (2025)	Graph Attention Networks (GAT) for IoT intrusion detection	Achieved **99% accuracy** in detecting IoT-based intrusions	High memory consumption and computational overhead in large-scale IoT environments
Ahmed et al., (2025)	Signature-based intrusion detection system	Improved precision and recall for attack detection,	Due to the reliance on signature-based approaches, there is

	that makes use of machine learning, deep learning, and fuzzy clustering	with **96.5% detection accuracy**	limited generalization to attacks that have not yet been seen.
Kumar et al., (2025)	Modified Graph Neural Network (GNN) with Explainable AI (XAI) for multi-class malware detection	Enhanced classification accuracy to improving malware categorization and explainability	Model complexity may hinder real-time detection capabilities
Wazid et al. (2025)	Explainable deep learning for IoT-enabled Intelligent Transportation Systems (ITS) malware detection	Achieved **99.7% detection accuracy**, improving threat detection in smart transportation	Model robustness against adversarial attacks remains a challenge

A varied range of machine algorithms can be used for analysis and their applications for vulnerability analysis and threat identification is performed and their performance are evaluated based on the parameters shown in the table (2).

Table 2. Matrix for Performance Analysis

Metric	**Description**
Accuracy	Percentage of correctly identified intrusion vs. benign traffic.
Precision	Proportion of actual intrusions among those predicted as intrusions. Reduces false positives.
Recall (Sensitivity)	Proportion of actual intrusions that were correctly identified. Reduces false negatives.
F1-Score	Harmonic mean of precision and recall, balancing both for imbalanced datasets.
Detection Time	Average time in milliseconds taken to detect an intrusion event.
Explainability Score	A subjective or model-derived score (e.g., SHAP values, rule extraction quality) on how well the model decisions can be understood by humans.

The different algorithms used are CNN, RNN, XGBoost, GNN, and EGL, their characterise such as, Local Reputation Field, Conventional Layers, Layer Stacking, Pooling Layers, Activation Functions, Fully Connected Layer and End-to-End Learning for the detection of security vulnerabilities is presented in the table (3).

Table 3. Matrix for Performance Analysis

Model	**Local Reputation field**	**Conventional Layers**	**Layer Stacking**	**Pooling Layer**	**Activation Functions**	**Fully Connected Layer**	**End-to-End Learning**
CNN	Capture spatial patterns and minimize computed cost and over fitting $y_{a,b} = \sum_{i=0}^{k=1} \sum_{j=0}^{k=1} w$ $x_a + i.b + j$ Detection of port scanning, DDOS, or brute force attacks	Multiple learnable filters in each layer $\mathrm{Conv}(k) = F \times k + b$ Detect SYN flood pattern	Hierarchical representation able to capture simple and abstract $h^{l+1} = \sigma(w^{(l)} \times h^{(l)} + b^{(l)}$ Able to detect advanced attacks such as APTs	Minimize sthe spatial dimension which reduces the number of parameters $h_{i,j} = \max\{x_{m,r}$ $m, n \in \mathrm{window}(i$ Make the IDS robust to noise and temporal shift	Learn complex patterns using non-linear activation function. ReLU, Leaky ReLU, ELU. $\sigma(x) = \max$ In anomaly detection	Towards the end a fully connected layer helps better prediction $y = \sigma(W_x + b)$ Intrusion detection	Gradient descent and back propagation and minimize the loss function. $L = -\sum_{i=1}^{N} y_i \log($ Detection of any evolving threats
RNN	Each RNN unit	Provide recent	Use of multi-	Focus on most	Functions	Maps hidden	Entire model

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

	processes input at time t with hidden state. $h_t = f(h_{t-1,}x_t$ It is able to detect the login failures or burst or increased traffic	features with RNN, LSTM, GRU $z_t = \mathrm{Conv1D}(x_t)$ HTTP based attacks, brute force	layer RNN $h_t^{(l)} = f(h_{t-1}^{(l)}, h_f^{l}$ Detect complex attack challenges such as APT	relevant time steps. $h_{pool} = \max_t(h_t)$ Privilege escalation attack	such as tanh, ReLU or sigmoid at each stage $\tanh(k = \frac{e^x -}{e^x +}$ $\mathrm{ReLU} = \max$ $\sigma(x) = \frac{1}{1 +}$ Anomaly detection	state to find prediction output $y = \mathrm{softma} + b)$ Classifies input sequence as malicious or benign.	trained using BPTT on labelled sequence data. $L = -\sum_{i=1}^{N} y_i lo$ Detection of new or evolving threats
XGBoost	Decision trees learn splits on local features selection Feature Im $= \sum_{t=1}^{T}\sum_{split\epsilon t}$ ΔLoss is the reduction in loss due	A group of decision tree series	Sequential boosting $\hat{y}_j = \sum_{t=1}^{T} f_t(x_i)$ Where F space regression trees	Aggregation via tree ensemble	Step-wise output at leave, no explicit activation like ReLU/tanh	Final output for labeling	Gradient boosting on structured loss $L\emptyset = \sum_{i=1}^{n} l(y_i, \hat{y}$ $+ \sum_{t=1}^{T} \Omega(f_t)$

Received: August 02, 2025

	to that split Detect the login attempt threshold	Capture nulti-feature attack patter	Privilege escalation	Anomaly detection	$f(x) = \sum_{j=1}^{J} V$ Decide input features are malicious or benign.	Intrusion detection: normal or DoS	Learning from log data
GNN	Neighborhood aggregation within k-map $h_v^{(k)} = AGG(N($ Detect localized attack behaviour	Message passing account, graph edges. $H^{(l+1)} = \sigma(\hat{A}H^{(l)}W^{(l)}$ Learning from structured dependencies such as attack trees.	Stacked layers for multi-hop dependency learning $h_v^{(k)} = GNNLay$ $\{h_u^{(k-1)} \| u\epsilon N$ Detect multistage or stealthy attacks across layers	Graph level readout or node subsampling $h_G = \frac{1}{\|v\|}\sum_{v\epsilon v} h_v$	Non-linearity each layer $\sigma(x) = ReL$ Complex mode of threat activities	Maps graph/node embedding to outputs. $\hat{y} = softma$ $+ b_f)$ Classifies each node/graph as benign or malicious.	Learn graph feature via back propagation $\sum_{v\epsilon v_{train}} Cross$ Detection of intrusion form topology
EGL	Graph convolution layers cumulative characteristics from neighbour	Graph convolution layers cumulative characteristics from neighbours in the graph,	Stacking multiple GNN layers allows learning from multi-hop	Graph pooling reduces the graph size by summarizing node informatio	Non-linear transformations applied after	After graph embedding, FC layers are used for classific	All components are trained together using a

	s in the graph, similar to how CNNs convolve over image patches. $h_v^{(k)} = \sigma(\sum_{u\epsilon N(v)} \alpha$ $)$ Detects local anomalies , peer behaviour	similar to how CNNs convolve over image patches. $h_v^{(k)} = \sigma(\sum_{u\epsilon N(v)} soft$ Captures contextual node behaviours	neighbourhoods. $h_v^l = GNNLay$ Models multi-hop attack behaviours	n, akin to max/avg pooling in CNNs. $S^{(l)} = softmax($ Graph-level intrusion summary	each graph layer. $\sigma(x) = ReL$ Enables complex pattern discrimination	ation or regression tasks. $\hat{y} = softma$ $+ b)$ Final classification (attack type, anomaly score)	unified loss function. $L = L_{Pred} + \lambda L_{expl}$ $L_{Pred} = Predictio$ $L_{expl} = Explanati$ λ =Regularization weight Improves accuracy and explainability of detections

The performance comparison using a test dataset from the public source is shown in the table (4) with their plot in the form of radar chart and graphical representation as in figure (4) and figure (5).

Table 4. Performance Comparison of EGRL with Other Models in Intrusion Detection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Detection Time (ms)	Explainability Score
EGL	96.9	95.7	97.4	97.1	120	9.1/10
CNN	94.1	92.3	91.9	92.0	160	3.4/10
RNN	91.9	90.0	90.7	90.3	180	2.8/10
XGBoost	94.6	94.5	94.8	94.5	140	4.7/10
GNN	95.6	93.7	96.1	94.9	130	5.2/10

The Radar Chart shows the comparison of the performance of Explainable Graphic Reinforcement Learning (EGRL) with other models across key metrics such as Accuracy (Blue), Precision (Green), Recall (Red) and F1 Score (Brown). The performance report indicates that EGRL consistently outperforms the other models across all parameters, indicating its effectiveness in intrusion detection tasks.

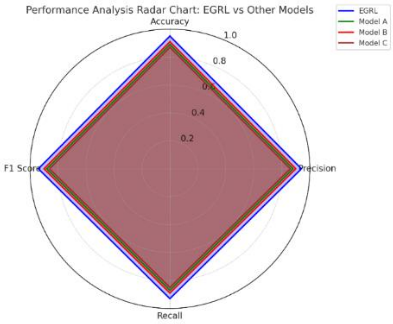

Figure 4. Radar Chart – to compare all metrics in a single graph

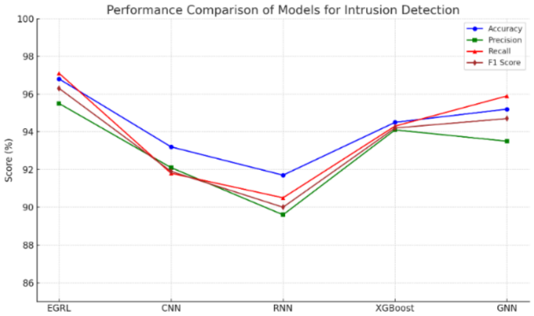

Figure 5. Performance Comparison of Models for Intrusion Detection

Conclusion

Though great advances have been achieved in graph-based network traffic monitoring, current IDS still battle scalability and real-time adaption. The conventional approaches fail to reflect the dynamic behavior of the network, therefore generating more false positives and reducing the detection accuracy. Although many models lack mechanisms for continuous retraining, which is required to properly handle evolving assault patterns, techniques for dynamic learning and adaptation have been studied. In addition, the integration of explainability in graph-based intrusion detection systems continues to be a difficulty. This is because the solutions that are now available prioritize detection accuracy over interpretability, which results in security decisions that are less visible. Furthermore, the evaluation of performance and explainability is frequently uneven and does not have defined benchmarks, which restricts the ability to compare different models simultaneously. Most AI-driven IDS fail to garner acceptability from cybersecurity professionals due to their black-box character, which is the reason why confidence and usability in these systems remain underexplored. Effectively addressing these research gaps will necessitate the development of novel graph-learning frameworks, upgrades to reinforcement learning, and improved interpretability methodologies for the purpose of enhancing user trust and system usability.

Future Scope

Future study ought to concentrate on enhancing graph-based network traffic monitoring through the development of scalable and real-time graph processing methodologies to manage extensive, dynamic environments. Improving dynamic learning and adaptation by self-learning, continually developing IDS models will enhance detection accuracy for emerging threats. Incorporating explainability techniques like SHAP, LIME, and attention mechanisms into graph-based learning would improve transparency, assisting cybersecurity professionals in comprehending AI-generated conclusions. Moreover, standardized evaluation frameworks must be established to systematically evaluate the trade-off between IDS performance and explainability, hence providing trustworthy benchmarking. Ultimately, cultivating trust and usability necessitates human-in-the-loop methodologies, wherein cybersecurity specialists engage with AI-driven Intrusion Detection Systems to authenticate warnings, enhance detection models, and augment reliability. By focusing on these aspects, future Intrusion Detection Systems will enhance adaptability, interpretability, and user-friendliness, hence assuring resilient real-time cybersecurity protections within intricate network infrastructures.

Acknowledgement

The authors extend the gratitude to Cyber Security Centre of excellence at Alliance University for supporting the completion of the research work.

References

1. Thapa, S., & Mailewa, A. (2020, April). The role of intrusion detection/prevention systems in modern computer networks: A review. In Conference: Midwest Instruction and Computing Symposium (MICS) (Vol. 53, pp. 1-14).
2. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity, 2(1), 1-22.
3. Mallick, M. A. I., & Nath, R. (2024). Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments. World Scientific News, 190(1), 1-69.
4. Mehta, G., Jayaram, V., Maruthavanan, D., Jayabalan, D., Parthi, A. G., Bidkar, D. M., ... & Veerapaneni, P. K. (2024). Emerging Cybersecurity Architectures and Methodologies for Modern Threat Landscapes. Journal ID, 9471, 1297.
5. Nie, M., Chen, D., & Wang, D. (2023). Reinforcement learning on graphs: A survey. IEEE Transactions on Emerging Topics in Computational Intelligence, 7(4), 1065-1082.
6. Devailly, F. X., Larocque, D., & Charlin, L. (2021). IG-RL: Inductive graph reinforcement learning for massive-scale traffic signal control. IEEE Transactions on Intelligent Transportation Systems, 23(7), 7496-7507.

7. Alwasel, B., Aldribi, A., Alreshoodi, M., Alsukayti, I. S., & Alsuhaibani, M. (2023). Leveraging graph-based representations to enhance machine learning performance in IIoT network security and attack detection. Applied Sciences, 13(13), 7774.
8. Ren, K., Zeng, Y., Zhong, Y., Sheng, B., & Zhang, Y. (2023). MAFSIDS: a reinforcement learning-based intrusion detection model for multi-agent feature selection networks. Journal of Big Data, 10(1), 137.
9. Sarker, I. H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. ICT Express.
10. Sayyed, T., Kodwani, S., Dodake, K., Adhayage, M., Solanki, R. K., & Bhaladhare, P. R. B. (2023). Intrusion Detection System. Int. J. of Aquatic Science, 14(1), 288-298.
11. Elrawy, M. F., Awad, A. I., & Hamed, H. F. (2018). Intrusion detection systems for IoT-based smart environments: a survey. Journal of Cloud Computing, 7(1), 1-20.
12. Kheddar, H. (2024). Transformers and large language models for efficient intrusion detection systems: A comprehensive survey. arXiv preprint arXiv:2408.07583.
13. Islam, R., Devnath, M. K., Samad, M. D., & Al Kadry, S. M. J. (2022). GGNB: Graph-based Gaussian naive Bayes intrusion detection system for CAN bus. Vehicular Communications, 33, 100442.
14. Caville, E., Lo, W. W., Layeghy, S., & Portmann, M. (2022). Anomal-E: A self-supervised network intrusion detection system based on graph neural networks. Knowledge-based systems, 258, 110030.
15. Dos Santos, R. R., Viegas, E. K., Santin, A. O., & Cogo, V. V. (2022). Reinforcement learning for intrusion detection: More model longness and fewer updates. IEEE Transactions on Network and Service Management, 20(2), 2040-2055.
16. Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., & Zomaya, A. Y. (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. Information Sciences, 639, 119000.
17. Le, T. T. H., Prihatno, A. T., Oktian, Y. E., Kang, H., & Kim, H. (2023). Exploring local explanation of practical industrial AI applications: A systematic literature review. Applied Sciences, 13(9), 5809.
18. Zhong, M., Lin, M., Zhang, C., & Xu, Z. (2024). A survey on graph neural networks for intrusion detection systems: methods, trends and challenges. Computers & Security, 103821.
19. Thang, V. V., & Pashchenko, F. F. (2019). Multistage System-Based Machine Learning Techniques for Intrusion Detection in WiFi Network. Journal of Computer Networks and Communications, 2019(1), 4708201.
20. Palmer, I., Rogers, E., & Mcfly, S. (2020). A Graph-Based Analysis of Industrial Control Systems Network Traffic.
21. Abou Daya, A., Salahuddin, M. A., Limam, N., & Boutaba, R. (2020). BotChase: Graph-based bot detection using machine learning. IEEE Transactions on Network and Service Management, 17(1), 15-29.

22. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. IEEE Access, 10, 112392-112415.
23. Baahmed, A. R. E. M., Andresini, G., Robardet, C., & Appice, A. (2023, September). Using graph neural networks for the detection and explanation of network intrusions. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 201-216). Cham: Springer Nature Switzerland.
24. Lo, W. W. (2023). Graph representation learning for cyberattack detection and forensics.
25. Kaya, K., Ak, E., Bas, S., Canberk, B., & Oguducu, S. G. (2024, June). X-CBA: Explainability Aided CatBoosted Anomal-E for Intrusion Detection System. In ICC 2024-IEEE International Conference on Communications (pp. 2288-2293). IEEE.
26. Adhikari, D., & Thapaliya, S. (2024). Explainable AI for Cyber Security: Interpretable Models for Malware Analysis and Network Intrusion Detection. NPRC Journal of Multidisciplinary Research, 1(9), 170-179.
27. Farrukh, Y. A., Wali, S., Khan, I., & Bastian, N. D. (2024). Xg-nid: Dual-modality network intrusion detection using a heterogeneous graph neural network and large language model. arXiv preprint arXiv:2408.16021.
28. Shokouhinejad, H., Razavi-Far, R., Mohammadian, H., Rabbani, M., Ansong, S., Higgins, G., & Ghorbani, A. A. (2025). Recent Advances in Malware Detection: Graph Learning and Explainability. arXiv preprint arXiv:2502.10556.
29. Kalutharage, C. S., Liu, X., & Chrysoulas, C. (2025). Neurosymbolic learning and domain knowledge-driven explainable AI for enhanced IoT network attack detection and response. Computers & Security, 151, 104318.
30. Ahanger, A. S., Khan, S. M., Masoodi, F., & Salau, A. O. (2025). Advanced intrusion detection in internet of things using graph attention networks. Scientific Reports, 15(1), 9831.
31. Ahmed, U., Nazir, M., Sarwar, A., Ali, T., Aggoune, E. H. M., Shahzad, T., & Khan, M. A. (2025). Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering. Scientific Reports, 15(1), 1726.
32. Kumar, S., Khot, V., Bhat, S., Ghare, A., & Kapadi, R. (2025). Multi-class Malware Detection using Modified GNN and Explainable AI. Frontiers of Innovation, 126.
33. Wazid, M., Singh, J., Pandey, C., Sherratt, R. S., Das, A. K., Giri, D., & Park, Y. (2025). Explainable Deep Learning-Enabled Malware Attack Detection for IoT-Enabled Intelligent Transportation Systems. IEEE Transactions on Intelligent Transportation Systems.