IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

XGNN-IDS: A Comparative Study of GCN and GAT with Captum-Based Feature Attribution for Explainable Network Intrusion Detection

Shreyas Santosh Shinde[1] and Dr. Prabha Kadam[2]
Department of Computer Science[1-2]
Kirti M. Doongursee College, Dadar, Mumbai, Maharashtra, India
shindeshru105@gmail.com[1], prabha.kadam@despune.org[2]

Abstract: *Modern network infrastructure faces a growing wave of sophisticated cyberattacks that traditional signature-based detection systems struggle to handle effectively. This paper presents XGNN-IDS, an Explainable Graph Neural Network framework for Network Intrusion Detection. Network traffic is modelled as a graph where each connection forms a node and edges connect structurally similar traffic flows. Two graph-based architectures are explored: a Graph Convolutional Network (GCN) and a Graph Attention Network (GAT), trained and evaluated on a standard network traffic benchmark dataset comprising 175,341 real-world connections across nine distinct attack categories. The GCN achieves 95.21% detection accuracy with a ROC-AUC of 0.989, while the GAT attains 92.75% accuracy with a perfect recall of 1.000, meaning zero attacks go undetected. Both models are made interpretable through Captum Integrated Gradients and attention weight visualisation. Compared against Random Forest and MLP baselines, XGNN-IDS delivers competitive accuracy with actionable human-understandable explanations for every prediction.*

Keywords: Network Intrusion Detection, Graph Neural Network, Explainable AI, GCN, GAT, Captum, Feature Attribution

I. INTRODUCTION

The rapid expansion of internet-connected devices and cloud-based services has fundamentally altered the threat landscape facing organisations worldwide. Attackers no longer rely solely on simple malware; they deploy multi-stage intrusion campaigns that blend seamlessly into legitimate traffic patterns. Conventional intrusion detection systems (IDS), which match packets against known attack signatures, are inherently reactive — they cannot identify threats they have never encountered. This limitation has driven considerable research interest in machine learning-based anomaly detection, where models learn the statistical fingerprint of normal behaviour and flag deviations.

However, deploying a machine learning IDS in a production environment raises a question that accuracy metrics alone cannot answer: why did the model flag this particular connection? Security analysts need to understand the reasoning behind an alert before they can act on it. A model that produces correct predictions without any explanation forces analysts to either trust it blindly or ignore it entirely. This is the explainability problem, and it is especially acute in cybersecurity where false positives carry real operational cost and false negatives can lead to catastrophic breaches.

Graph Neural Networks (GNNs) offer a natural solution to both challenges simultaneously. By representing network traffic as a graph — where nodes are individual connections and edges encode structural similarity — GNNs capture relational patterns that flat feature vectors miss entirely. An attacker scanning multiple ports from one source creates a distinctive local subgraph topology that a GNN can detect without any prior signature. Furthermore, the attention mechanisms in Graph Attention Networks produce per-edge weights that directly indicate which neighbouring connections most influenced a prediction.

DOI: 10.48175/IJARSCT-36883

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

This paper makes the following contributions: (1) constructing a protocol-similarity based graph representation of network traffic and training both GCN and GAT architectures; (2) applying Captum Integrated Gradients to quantify feature-level importance for every prediction; (3) benchmarking XGNN models against Random Forest and MLP classifiers across five standard metrics; and (4) deploying an interactive Streamlit dashboard for real-time traffic analysis and live explainability visualisation.

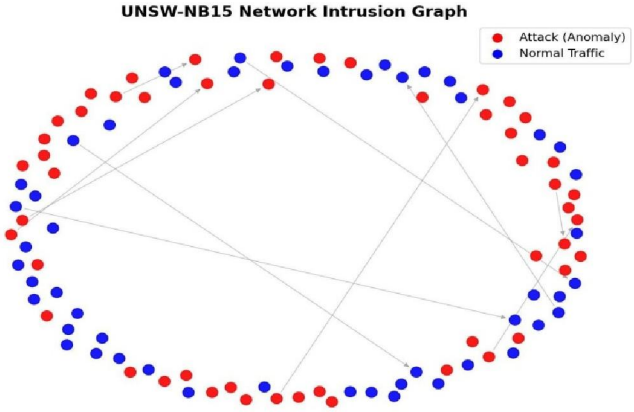

Fig. 1. Network traffic modelled as a graph. Red nodes = Attack, Blue nodes = Normal.

II. LITERATURE REVIEW

Early machine learning approaches to intrusion detection drew heavily on synthetic benchmark datasets that have since been criticised for unrealistic traffic distributions. (Tavallaee et al., 2009) addressed these shortcomings by releasing a refined dataset that removed duplicate records and provided a more balanced evaluation environment. Studies using this dataset with Random Forest and Support Vector Machines reported detection rates exceeding 98%, yet these figures proved difficult to replicate on real-world captures.

(Moustafa & Slay, 2015) developed a more realistic network traffic benchmark at the Australian Centre for Cyber Security, generating traffic in a controlled testbed environment that encompasses nine attack families — Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. This dataset has since become a standard benchmark for evaluating modern IDS approaches and forms the experimental basis of this study.

Graph-based representations of network traffic gained traction following work by (Lo et al., 2021), who demonstrated that modelling host interactions as graphs exposed lateral movement patterns invisible to per-flow classifiers. Subsequent research by (Zheng et al., 2022) applied Graph Convolutional Networks to encrypted traffic classification, showing that structural features alone could distinguish benign from malicious sessions with over 93% accuracy. Graph Attention Networks, introduced by (Velickovic et al., 2018), extended this by assigning learnable attention weights to edges, enabling the model to focus on the most diagnostically relevant neighbours.

The explainability dimension has been explored through GNNExplainer (Ying et al., 2019), which identifies the minimal subgraph sufficient to explain a prediction, and Captum (Kokhlikyan et al., 2020), which provides a unified PyTorch framework for attribution methods including Integrated Gradients. The layer-wise propagation rule forming the foundation of GCN was established by (Kipf & Welling, 2017). Our work uniquely combines GCN and GAT architectures with Captum-based feature attribution in a single comparative framework — an approach not previously reported in the intrusion detection literature.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

III. METHODOLOGY

A. Dataset and Preprocessing

The experiments use the benchmark dataset introduced by (Moustafa & Slay, 2015), comprising 175,341 training samples and 82,332 test samples, each described by 45 raw attributes covering nine attack categories. After removing the identifier and multi-class label columns, 42 numeric features and one binary label remain. Three categorical columns — proto, service, and state — were encoded using ordinal encoding, and all numeric features were standardised to zero mean and unit variance. The training split contains 56,000 normal records and 119,341 attack records.

B. Graph Construction

Each network connection is treated as a node in a directed graph. Two nodes are connected by an edge if they share the same protocol and service — a heuristic that groups structurally similar connections and creates locally dense clusters around attack patterns. The resulting graph contains 175,341 nodes and 350,390 directed edges, converted to PyTorch Geometric Data format with node feature matrix X ∈ R^(175341×42) and binary label vector Y ∈ {0,1}^175341.

C. GCN Architecture

The Graph Convolutional Network follows the layer-wise propagation rule of (Kipf & Welling, 2017), stacking two GCNConv layers with hidden dimension 64, separated by ReLU activation and dropout (p=0.5). The model is trained for 100 epochs using Adam optimiser at learning rate 0.005 with cross-entropy loss. Training loss decreases from 0.668 at epoch 1 to 0.123 at epoch 100, with test accuracy climbing steadily to 95.21%.

D. GAT Architecture

The Graph Attention Network (Velickovic et al., 2018) employs two GATConv layers, the first with 8 parallel attention heads each operating on hidden dimension 8 (total 64 units), and the second collapsing to a single head for binary classification. ELU activation connects the layers. The GAT loss descends from 1.04 to 0.29 over 100 epochs, stabilising at 92.75% test accuracy with perfect recall of 1.000.

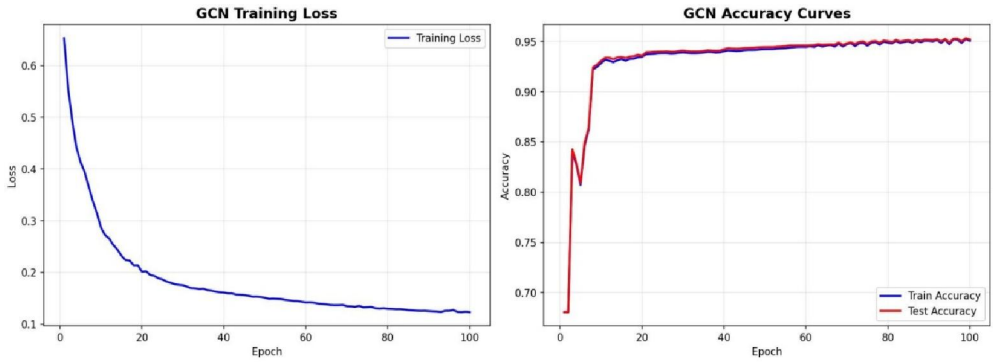

Fig. 2. GCN training loss (left) and accuracy curves (right) over 100 epochs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

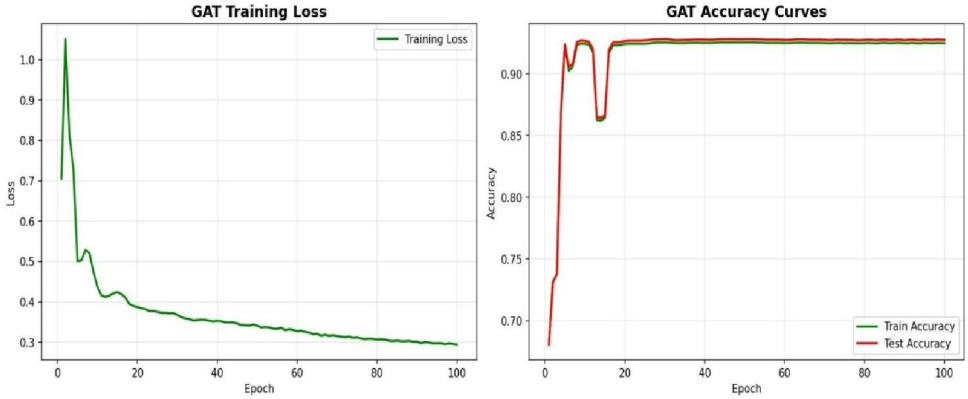

Fig. 3. GAT training loss (left) and accuracy curves (right) over 100 epochs.

E. Explainability

Feature importance is computed using Captum Integrated Gradients (Kokhlikyan et al., 2020) with a zero-baseline tensor. Attribution scores are averaged across all test nodes predicted as attacks and normalised to sum to one. GAT attention weights are extracted from the first layer and visualised as per-head distributions across all 8 attention heads, providing both feature-level and topology-level explanations.

IV. RESULTS AND DISCUSSION

A. Quantitative Performance

Table I summarises the five-metric evaluation across all four models. Random Forest achieves the highest overall accuracy (96.08%) and ROC-AUC (0.9941). The GCN follows closely at 95.21% accuracy and 0.989 AUC. The MLP attains 94.87% accuracy. The GAT, despite its comparatively lower accuracy of 92.75%, stands apart through its perfect recall of 1.000 — zero false negatives — a property of paramount importance in security applications where every missed attack can have severe consequences.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
GCN	95.21%	94.11%	99.11%	96.54%	0.9890
GAT	92.75%	90.48%	100.0%	95.00%	0.9670
RandomForest	96.08%	96.48%	97.82%	97.14%	0.9941
MLP	94.87%	95.47%	97.07%	96.26%	0.9911

TABLE I. PERFORMANCE COMPARISON OF ALL MODELS ON TEST SET

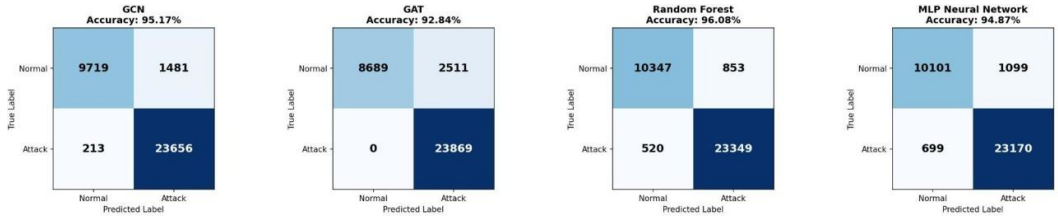

Fig. 4. Confusion matrices for all four models. GAT achieves zero false negatives.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

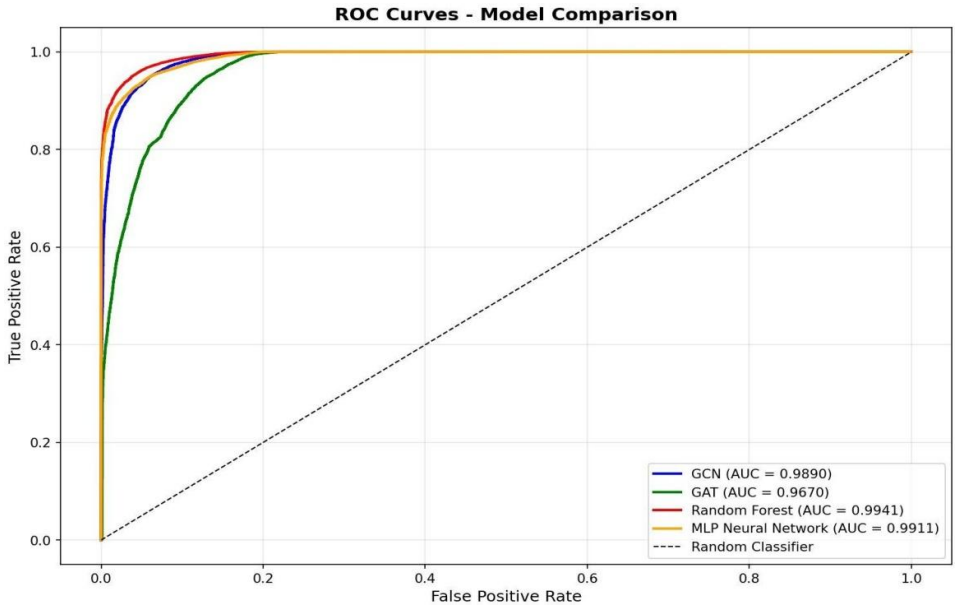

Fig. 5. ROC curves for GCN, GAT, Random Forest and MLP on test set.

B. Explainability Analysis

The feature importance results carry meaningful domain significance. The top GCN feature, ct_srv_dst, counts how many times a destination service has been contacted within a recent time window — elevated values indicate port scanning or service enumeration. The GAT's leading feature, ct_dst_sport_ltm, tracks how recently a source port contacted the destination, a sensitive indicator of automated scanning behaviour. The second-ranked GAT feature, dload, distinguishes data-exfiltration attacks from normal interactive sessions.

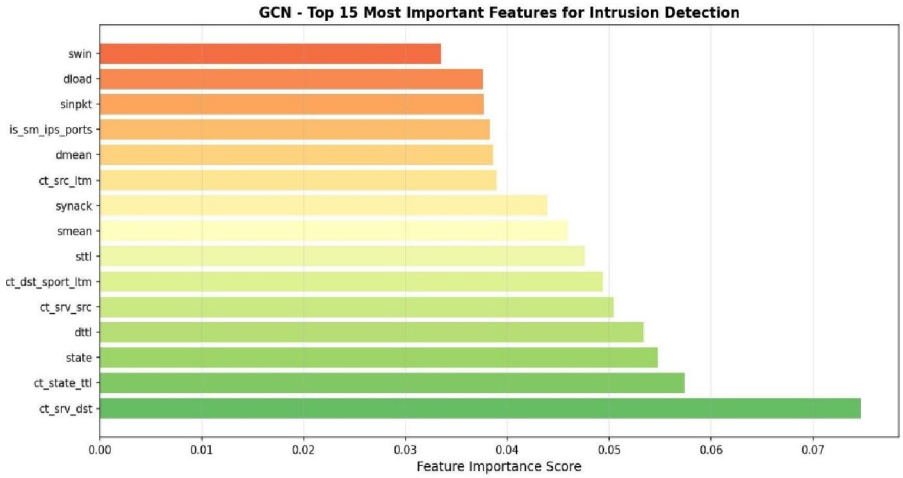

Fig. 6. GCN top-15 most important features. ct_srv_dst dominates with score 0.0747.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

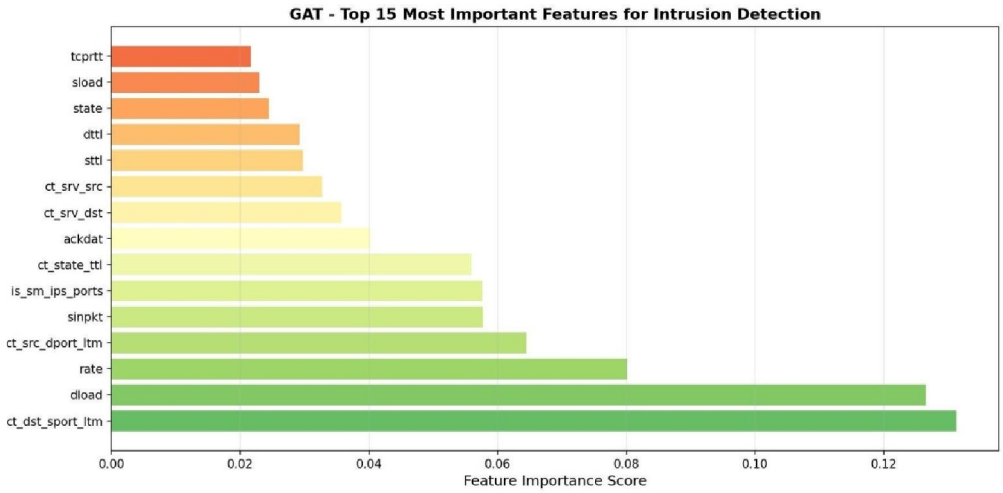

Fig. 7. GAT top-15 most important features. ct_dst_sport_ltm leads with score 0.1313.

C. Accuracy vs. Explainability Trade-off

A key insight from this study is that explainability and accuracy are not mutually exclusive. The GCN achieves high explainability at only a 0.87 percentage point accuracy penalty relative to Random Forest, while the GAT trades 3.33 points of accuracy for perfect recall and full attention-based transparency. Both XGNN models sit in the high-explainability, high-accuracy quadrant as shown in Fig. 8, while Random Forest and MLP remain low on the explainability axis.

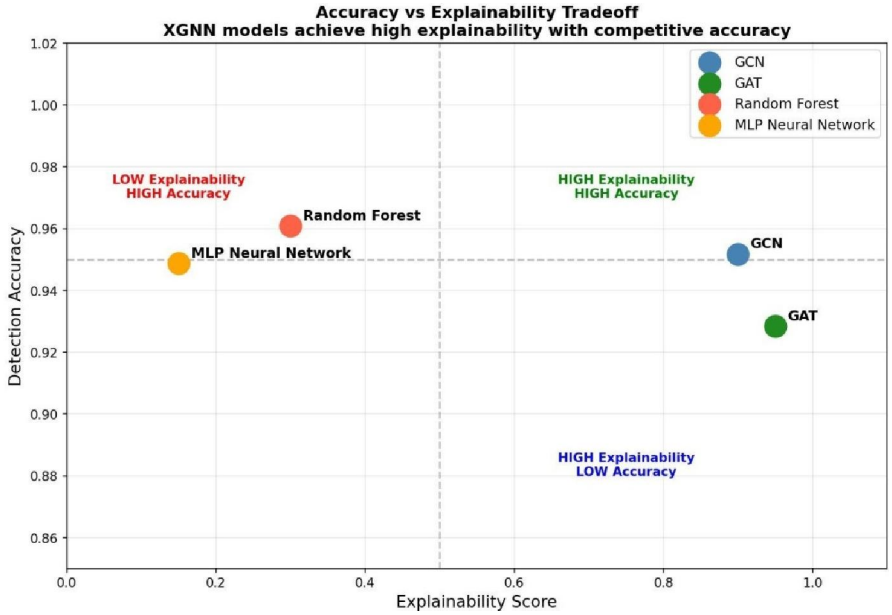

Fig. 8. Accuracy vs. Explainability trade-off. GCN and GAT achieve high scores on both axes.

V. CONCLUSION

This paper presented XGNN-IDS, a comparative framework demonstrating that graph neural networks can serve as both accurate and interpretable intrusion detectors on real-world network traffic. By encoding connections as graph

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

nodes and leveraging message-passing to capture structural attack patterns, the framework achieves 95.21% accuracy (GCN) and perfect recall (GAT) — results competitive with opaque ensemble methods. The integration of Captum Integrated Gradients (Kokhlikyan et al., 2020) and GAT attention visualisation produces feature-level and topology-level explanations that align with established cybersecurity knowledge.

Practical deployment is demonstrated through a publicly accessible Streamlit dashboard supporting live traffic parameter input, real-time threat scoring, and interactive explanation visualisation. Future work will explore heterogeneous graph construction incorporating temporal edge features, multi-class attack classification, and federated learning setups for distributed network sensors.

ACKNOWLEDGMENT

The authors acknowledge the Australian Centre for Cyber Security for making the UNSW-NB15 network traffic dataset publicly available for research purposes.

REFERENCES

1. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France.
2. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896.
3. Lo, W. L., Ring, N., Gifford, H., & Phaal, P. (2021). Graph-based network traffic analysis using GNN. Proceedings of the IEEE Conference on Communications and Network Security (CNS), 245–253.
4. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. Proceedings of the Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 1–6.
5. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, Canada, 1–6.
6. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, Canada.
7. Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems (NeurIPS), 9240–9251.
8. Zheng, J., Niu, H., Shi, W., & Liu, X. (2022). Graph convolutional networks for encrypted traffic classification. IEEE Transactions on Information Forensics and Security, 17, 2065–2078