Enhancing unknown attack detection for GNN-based NIDS through Open Set Recognition Approach

Thanh-Tung Nguyen[1], Minho Park[2]
[1]Department of Information Communication Convergence Technology,
Soongsil University, Seoul 156-743, South Korea
[2]School of Electronic Engineering, Soongsil University, Seoul 156-743, South Korea
tung2303.grad@gmail.com, mhp@ssu.ac.kr

***Abstract*—Network Intrusion Detection Systems (NIDS) play a crucial role in ensuring the security and integrity of computer networks. However, traditional NIDS models often struggle to identify unknown or emerging attacks, as they are typically trained on predefined attack patterns and normal behaviors. To address this limitation, this study integrates the principles of Open Set Recognition (OSR) into a Graph Neural Network (GNN)-based Intrusion Detection System (IDS). The proposed framework enables the model to not only detect known intrusions effectively but also to recognize unseen or novel attack instances by classifying them as unknown. By leveraging the structural learning capability of GNNs and the open-set decision boundary of OSR, the system enhances generalization and adaptability against evolving cyber threats. This approach aims to improve the robustness of NIDS in dynamic and continuously changing network environments. Results on the CIC-IDS2017 and UNSW-NB15 benchmarks indicate that the proposed method significantly improves Open Set Recognition performance and mitigates the issue of unknown attack detection.**

***Index Terms*—Intrusion Detection System (IDS), Graph Neural Network (GNN), Machine Learning, Flow-based Characteristic, Open Set Recognition**

I. Introduction

With the rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML), their applications in network security have gained increasing attention, particularly in the development of intelligent NIDS. Among various AI-driven approaches, GNN [1] have recently emerged as a powerful tool for modeling complex network structures and capturing relational dependencies between network entities, thereby improving the capability of intrusion detection systems to identify sophisticated cyberattacks. However, despite their promising performance, most existing IDS models—including those based on GNNs—suffer from a critical limitation: their inability to effectively recognize new or unknown attack types. In real-world environments, cyberattacks continuously evolve, with novel intrusion patterns appearing daily. Since these unseen attacks are not included during the training phase, the model tends to misclassify them into known categories, leading to incorrect decisions and reduced reliability. This highlights the urgent need for a mechanism that enables the system to detect and handle unknown classes, ensuring robust and adaptive protection against ever-changing network threats.

Recent studies have explored Transfer Learning (TL) and Continual Learning (CL) to improve the adaptability of Intrusion Detection Systems (IDS) to unseen or evolving attacks. These methods aim to help models retain previous knowledge while learning new patterns. However, both approaches have limitations in open-world network environments. In TL [2] [3], performance heavily depends on the similarity between the source and target domains. When new attacks differ significantly from the training data, transferred knowledge becomes unreliable. Meanwhile, CL [4] [5] focuses on updating models sequentially but still assumes that all target classes are known beforehand. Consequently, both methods struggle to handle truly unknown attacks, emphasizing the need for a more flexible framework capable of identifying unseen classes during detection.

Traditional Intrusion Detection Systems are generally developed under a closed-set assumption, where models are trained and tested on a fixed set of known attack types. While effective in controlled settings, this assumption fails in real-world environments where new attacks emerge constantly, causing the model to misclassify unseen patterns as known ones and leading to unreliable detection results. To overcome this issue, the concept of Open Set Recognition (OSR) [6] has been introduced. OSR enables models to correctly classify known classes while identifying unseen samples as unknown. Existing OSR approaches—such as threshold-based rejection [7], distance-based methods [8], and probabilistic modeling—aim to expand the decision boundary to handle unfamiliar data. However, most OSR research has focused on image or text domains, with limited exploration in Graph Neural Network (GNN)-based Intrusion Detection Systems. This work presents one of the first attempts to integrate OSR into a GNN-based IDS. Leveraging the message-passing mechanism of GNNs, which learns from node features and neighborhood relationships, our approach captures complex network dependencies, making it well-suited for open-set intrusion detection in dynamic network environments.

In this paper, we propose a GNN-based Intrusion De-

tection System that integrates the OpenMax method from Open Set Recognition (OSR) into a Graph Convolutional Network (GCN). The goal is to enable the model to classify known attacks accurately while effectively detecting unknown intrusions. OpenMax is chosen because it extends the traditional softmax layer by modeling class activation distributions using the Weibull function, allowing the model to estimate the likelihood that a sample belongs to an unknown class. This statistical approach provides a more flexible decision boundary compared to conventional classifiers. By combining the relational learning capability of GCNs with the open-set awareness of OpenMax, the proposed framework improves both accuracy and robustness, offering better adaptability to emerging and previously unseen network attacks.

We summarize our contributions as follows:

- We propose an open-set NIDS that integrates the Weibull-based probabilistic mechanism into a GCN to enhance the capability of detecting both known and unknown attacks.
- We evaluate the proposed model on two benchmark datasets, CIC-IDS2017 and UNSW-NB15, and demonstrate that our approach significantly improves Open Set Recognition performance and enhances the robustness of GNN-based IDS in dynamic network environments.

The rest of this paper is organized as follows. In Section II, we introduce the proposed methodology; Section III presents the experimental evaluation; and Section IV concludes the study.

II. Proposed Method

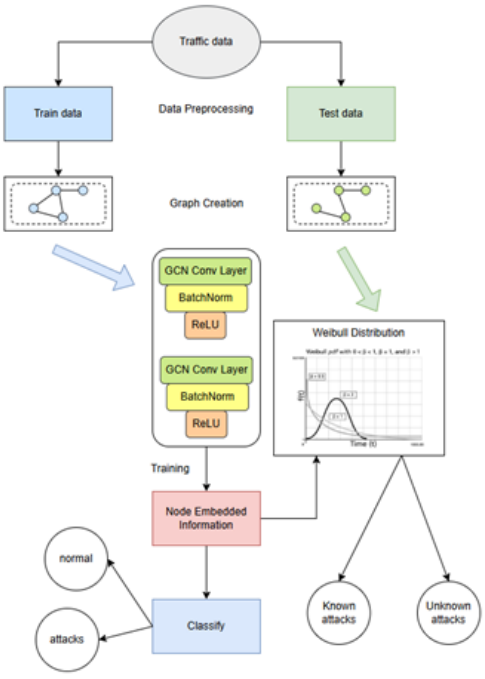

Fig. 1: Diagram of the proposed model.

The proposed framework for NIDS using a GCN-based Open Set Recognition approach can be explained through several sequential stages, as illustrated in the Fig. 1.

First, traffic data are collected and subjected to a data preprocessing phase. In this step, the raw network traffic is cleaned, normalized, and structured to ensure data consistency. The data are then divided into training and testing subsets to support model development and evaluation.

Next, both the training and testing data are transformed into a graph representation. In this graph, nodes represent entities such as hosts or network flows, while edges denote communication or relationships between them. This graph-based transformation captures the structural dependencies inherent in network interactions, which is essential for applying GCN.

After graph creation, the GCN model processes the graph data through multiple layers. Each layer consists of a graph convolutional layer, followed by batch normalization and a ReLU activation function. This combination enables the model to extract high-level node embeddings by aggregating information from neighboring nodes, effectively learning both local and global network features.

During the training phase, the node embedding information generated by the GCN. presented in equation 1, is used for classification. The model learns to distinguish between normal and attack instances based on learned graph representations.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{1}$$

where:

- $H^{(l)}$ is the feature matrix at layer l (with $H^{(0)} = X$, the input features),
- $\tilde{A} = A + I$ is the adjacency matrix of the graph with added self-loops,
- $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$,
- $W^{(l)}$ is the trainable weight matrix at layer l,
- $\sigma(\cdot)$ is a non-linear activation function, e.g., ReLU.

To handle unseen or unknown attacks, the proposed method incorporates the Weibull distribution, as presented in equation 2. After training, the model applies the Weibull fitting on the distance scores between known classes in the embedding space. This allows the system to estimate a probability boundary for identifying known and unknown attack samples. As a result, the model not only performs traditional classification but also extends its ability to detect previously unseen intrusions, enhancing the robustness of the NIDS.

$$F(d;\lambda,k) = 1 - e^{-(d/\lambda)^k} \tag{2}$$

where $\lambda > 0$ is the scale parameter and $k > 0$ is the shape parameter. These parameters are estimated using the distances of the extreme activation vectors (the largest distances) for each class.

III. Result and Discussion

Experiments were performed on two benchmark datasets, CIC-IDS2017 and UNSW-NB15, comprising 26,554 and

56,000 network flows, respectively. For each dataset, the data was partitioned such that 75% was allocated for training and the remaining 25% for testing.

To examine the misclassifications of unknown attacks, experiments were conducted on the CIC-IDS2017 and UNSW-NB15 datasets. In each experiment, several classes were deliberately excluded from the training set and included only in the testing set. This setup caused traditional models to misclassify these unseen classes with high confidence, as illustrated in Fig. 2. Following evaluation, the model achieved poor performance, with an accuracy of 70.45% on CIC-IDS2017 and 56.73% on UNSW-NB15, highlighting the limitations of conventional GCN models when confronted with previously unseen attacks.

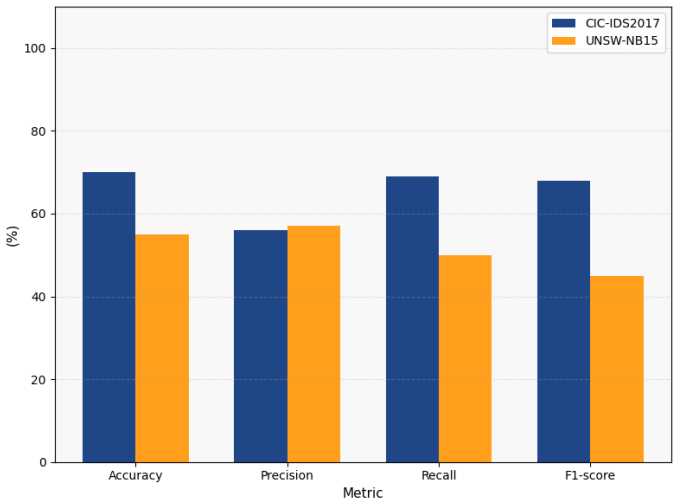

Fig. 2: Model misclassified on unknown attacks.

The performance of the proposed model is presented in Fig. 3. Upon evaluation, the model achieved prediction accuracies of 91.56% on the CIC-IDS2017 test set and 93.73% on UNSW-NB15. As a result, the misclassifications rate for unknown attacks was substantially reduced, demonstrating the model's effectiveness. Furthermore, performance improvements are notable compared to the results shown in Fig. 2. These findings indicate that the proposed framework significantly enhances the capability of GNN-based NIDS models to detect and recognize previously unseen attacks.

IV. Conclusion

In this work, we propose a novel framework that integrates Weibull distribution methods with a GNN-based IDS model incorporating Open Set Recognition (OSR), aiming to enhance its capability to detect and recognize unknown attacks. By modeling the tail of class-wise activation distributions with the Weibull approach, our framework effectively distinguishes between known and unseen attack patterns. Experimental results on benchmark datasets demonstrate the robustness and effectiveness of the proposed method. In future work, we plan to further optimize the model using advanced sample selection strategies and conduct more comprehensive evaluations across diverse datasets.

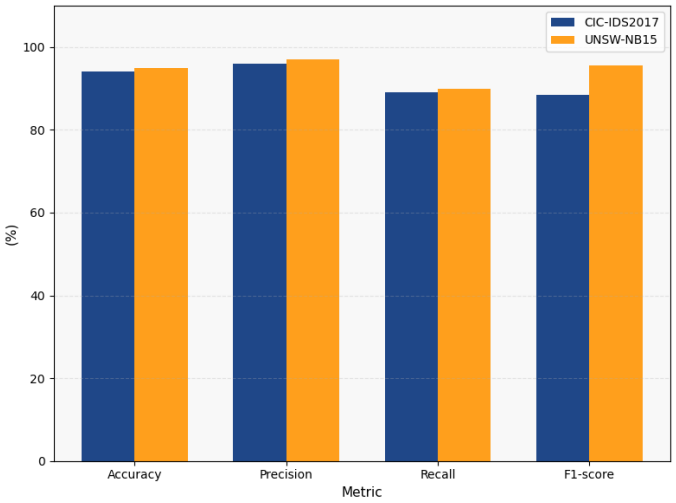

Fig. 3: Proposed method performance.

Acknowledgment

This work was jointly supported by the National Research Foundation of Korea (NRF) via a grant provided by the Korea government (MSIT) (grant No. NRF-2023R1A2C1005461) and by the MSIT (Ministry of Science and ICT), Korea, under the Convergence security core talent training business support program (IITP-2024-RS-2024-00426853) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

References

[1] T. Bilot, N. E. Madhoun, K. A. Agha, and A. Zouaoui, "Graph neural networks for intrusion detection: A survey," *IEEE Access*, vol. 11, pp. 49114–49139, 2023.

[2] S. Ola-Obaado and M. A. Suleiman, "Anomaly-based network intrusion detection using transfer learning," in *2023 2nd International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS)*, vol. 1, pp. 1–5, 2023.

[3] H.-M. Chuang and L.-J. Ye, "Applying transfer learning approaches for intrusion detection in software-defined networking," *Sustainability*, vol. 15, no. 12, 2023.

[4] T.-T. Nguyen and M. Park, "El-gnn: A continual-learning-based graph neural network for task-incremental intrusion detection systems," *Electronics*, vol. 14, no. 14, 2025.

[5] S. kumar Amalapuram, S. S. Channappayya, and B. Tamma, "Augmented memory replay-based continual learning approaches for network intrusion detection," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[6] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3614–3631, 2020.

[7] S. Cruz, R. Rabinowitz, M. Günther, and T. E. Boult, "Operational open-set recognition and postmax refinement," in *European Conference on Computer Vision*, pp. 475–492, Springer, 2024.

[8] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.