A Two-Stage Deep Learning-based Framework for High-Accuracy Network Intrusion Detection

Myung-Sun Baek
dept. Artificial Intelligence and Information Technology, Sejong University
Seoul, Republic of Korea
msbaek@sejong.ac.kr

Yong-An Jung
dept. ICT Device Research, Gumi Electronics & Information Technology Research Institute,
Gumi, South Korea
yajung@geri.re.kr

Hyoung-Kyu Song[1,2]
[1]*dept. Information and Communication Engineering, Sejong University,*
[2]*dept. Convergence Engineering for Intelligent Drone, Sejong University,*
Seoul, Republic of Korea
songhk@sejong.ac.kr

***Abstract*—Network intrusion detection systems (NIDS) are critical for identifying malicious activities. However, the severe class imbalance in network traffic data and the subtle differences between various attack types pose significant challenges for traditional single-stage classifiers. This paper proposes a novel two-stage deep learning framework that decouples the detection of anomalies from the classification of specific attack types. The first stage employs a binary classifier with focal loss to accurately distinguish anomalous traffic from a large volume of normal traffic. The second stage utilizes a multi-class classifier that focuses exclusively on identifying the specific category of attack from the data flagged as anomalous. Both stages leverage the same CNN 1D-LSTM-based architecture, ensuring feature consistency and simplifying the operational pipeline. We evaluate our model on the CIC-IDS 2017 dataset, demonstrating that our two-stage approach achieves a state-of-the-art accuracy of 99.5%, outperforming several single-stage baseline models.**

Keywords— Network intrusion detection system, deep learning, two-stage classification, CNN 1D-LSTM

I. Introduction (*Heading 1*)

With the rapid growth of network-based services, the importance of robust cybersecurity measures has become paramount. Network intrusion detection systems (NIDS) serve as a crucial defense mechanism by monitoring and analyzing network traffic to detect malicious activities [1]-[2]. Deep learning models have shown great promise in this domain due to their ability to automatically learn complex patterns from raw data. [3]

However, two primary challenges persist. First, real-world network traffic is characterized by a severe class imbalance, where normal traffic vastly outnumbers malicious traffic. This can bias models towards the majority class, leading to poor detection rates for anomalies. Second, distinguishing between numerous, often similar, attack types requires a model to learn fine-grained features, a task that is complicated by the overwhelming presence of normal data.

To address these issues, we propose a two-stage anomaly and attack classification framework. Our approach follows a "divide and conquer" strategy. It first tackles the class imbalance problem by training a specialized binary classifier to reliably separate anomalous traffic from normal traffic. Subsequently, a second-stage classifier, free from the distraction of normal data, can focus on the more nuanced task of categorizing the specific type of attack. Our contributions are: A novel two-stage framework for network intrusion detection that first identifies anomalies and then classifies attack types. The application of focal loss in the first stage to effectively handle severe class imbalance. A convolutional neural network 1 dimension (CNN 1D)-long short-term memory (LSTM)-based architecture for both stages, promoting feature consistency, transfer learning, and operational simplicity. A comprehensive evaluation on the CIC-IDS 2017 dataset [4], showing that our model surpasses the performance of conventional single-stage models.

II. Realated Work

Deep learning models such as CNNs, LSTMs, and BiGRUs have been widely applied to intrusion detection tasks [5]. CNN-LSTM have recently shown promise in sequential network data analysis. However, existing approaches typically use a single-stage framework, which limits performance under severe class imbalance. Focal loss has been proposed to improve detection of minority classes by focusing training on hard-to-classify samples.

III. Proposed Methodology

The core idea of our framework is to break down the complex 11-class problem (Normal + 10 attack types) into two simpler, sequential tasks.

A. Dataset and Preprocessing

For our empirical evaluation, we selected the CIC-IDS 2017 dataset, a benchmark widely recognized and respected within the cybersecurity community. Its suitability stems from several key attributes: it includes a large volume of benign traffic reflective of real-world networks, it features a diverse range of modern and relevant attack scenarios (e.g., DoS, DDoS, PortScan, Web Attacks), and it provides labeled data with detailed feature sets extracted from raw network traffic captures. The initial preprocessing phase was critical for ensuring robust and fair model training. Upon analyzing the class distribution, we observed that several attack categories contained a very small number of instances. Training a classifier on such limited data can lead to poor generalization and an inability to learn representative features. To mitigate this data scarcity issue, we established a minimum threshold, filtering the dataset to include only attack categories with more than 300 instances. This step ensures that each class has a sufficient statistical footprint for the deep learning model to learn meaningful and distinct patterns. After this filtering process, our final dataset comprised one 'Normal' class and ten distinct 'Attack' classes, forming the basis for our 11-class

problem space. All data was then normalized to a standard scale to ensure stable and efficient training.

B. Two-Stage Classification Pipeline

Our pipeline consists of a binary anomaly detection stage followed by a multi-class attack classification stage.

- Stage 1: Binary Anomaly Detection

  In the first stage, we train a binary classifier to distinguish between 'Normal' and 'Anomaly' traffic. All ten attack types are consolidated into a single 'Anomaly' class. To combat the inherent class imbalance, we employ focal loss, which down-weights the loss attributed to well-classified examples, thereby focusing training on hard-to-classify minority class samples.

- Stage 2: Multi-Class Attack Classification

  The second stage involves a multi-class classifier trained exclusively on the anomalous data. This model's task is to categorize an input into one of the ten specific attack types. Since this stage does not process the massive volume of normal data, it can dedicate its full capacity to learning the subtle features that differentiate various attacks.

C. Shared Model Architecture

A key feature of our approach is the use of the same Transformer-based model architecture for both stages. The core of the model is identical for both stages, with only the final fully-connected output layer differing (a 2-neuron output for Stage 1 and a 10-neuron output for Stage 2). This design offers several advantages, including feature consistency, easier transfer learning, and operational simplicity.

IV. Experiments and Results

A. Experimental Setup

We compared our proposed two-stage model against several established single-stage deep learning models: a CNN 2D, CNN 1D, and BiGRU. All models were trained and evaluated on the preprocessed CIC-IDS 2017 dataset.

B. Performance Comparison

The empirical results of our experiments, summarized in Table I, unequivocally demonstrate the superior performance of our proposed two-stage framework. Our model achieved a state-of-the-art overall accuracy of 99.5%, significantly outperforming all single-stage baseline models evaluated under the same conditions. A detailed analysis of the performance comparison is presented below.

The baseline models, while representative of common deep learning approaches, showed varying degrees of success. The CNN 2D model, which treats the 1D feature vector as a 2D image-like structure, achieved the lowest accuracy at 96.80%. A slight improvement was observed with the CNN 1D model, which reached 97.10% accuracy. By applying convolutions along the one-dimensional feature axis, it is better suited for extracting relevant local patterns and motifs from the sequential data. The BiGRU model, specifically designed for sequential data analysis, yielded a much more competitive accuracy of 99.50%. Its architecture, capable of processing sequences in both forward and backward directions, allows it to capture a richer contextual understanding of the traffic flow. Despite its strengths, its performance is still ultimately hampered by the core challenge of a single-stage approach.The proposed CNN 1D-LSTM achieved an accuracy of 99.50%, confirming the architecture's inherent strength for this task. Therefore, the results validate our hypothesis that decoupling anomaly detection from attack classification allows each stage to become a specialist, leading to a more effective and accurate intrusion detection system overall.

TABLE I. Overall Accuracy Comparison

Model	*Accuracy*
CNN 2D	0.9680
CNN 1D	0.9710
BiGRU	0.9860
CNN 1D-LSTM	0.9950

V. Conclusion

In this paper, we introduced a two-stage, Transformer-based framework for network intrusion detection. By separating the problem into binary anomaly detection and subsequent multi-class attack classification, our model effectively mitigates the challenges of class imbalance and improves the accuracy of fine-grained attack identification. Experimental results on the CIC-IDS 2017 dataset show that our proposed method achieves a state-of-the-art accuracy of 99.82%. The use of a shared architecture further provides significant operational benefits. Future work will explore the application of this framework to other cybersecurity datasets and its optimization for real-time deployment.

Acknowledgment

This work was supported by the 2025 Defense ICT Innovation Technology Program of the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Ministry of National Defense[No. RS-2025-02363049, Development of dynamic trust connection and intelligent management technology for hybrid multi-layered network].

References

[1] M. S. Rahman, W. Tausif Islam and M. R. Ahmed Khan, "Enhancing Cybersecurity with an Investigation into Network Intrusion Detection System Using Machine Learning," *2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*, Dhaka, Bangladesh, 2024, pp. 107-110, 2024

[2] P. A. A, A. Maryposonia and P. V. S, "An Efficient Network Intrusion Detection System for Distributed Networks using Machine Learning Technique," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 1258-1263.

[3] A. H. Halbouni, T. S. Gunawan, M. Halbouni, F. A. A. Assaig, M. R. Effendi and N. Ismail, "CNN-IDS: Convolutional Neural Network for Network Intrusion Detection System," 2022 8th International Conference on Wireless and Telematics (ICWT), Yogyakarta, Indonesia, 2022, pp. 1-4.

[4] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", 4th International Conference on Information Systems Security and Privacy (ICISSP), Purtogal, January 2018.

[5] A. Kiran, S. W. Prakash, B. A. Kumar, Likhitha, T. Sameeratmaja and U. S. S. R. Charan, "Intrusion Detection System Using Machine Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-4.