LSTM-Based Network Intrusion Detection System and Solving Data Imbalance Problem through GAN

HAE-WON JEONG[1], HYEONG-GEON KIM[2], YOON-HO CHOI[3]
[1,2]*School of Computer Science and Engineering*, *Pusan National University*
Busan 46241, Republic of Korea
Email: speedheawon@pusan.ac.kr[1], qddd2000@pusan.ac.kr[2]
[3]*Professor, School of Computer Science and Engineering*, *Pusan National University*
Busan 46241, Republic of Korea
Email: yhchoi@pusan.ac.kr

***Abstract*—With the increasing sophistication and variety of cyberattacks in network environments, traditional rule-based intrusion detection systems have proven insufficient to address advanced threats such as Advanced Persistent Threats (APTs). This study presents an LSTM-based network intrusion detection model that incorporates GAN-based oversampling to address the class imbalance issue commonly found in network traffic data sets. By generating synthetic attack samples, the proposed model aims to enhance the anomaly detection performance. Comparative experiments with alternative approaches, including SMOTE and One-Class SVM, demonstrate the strengths and weaknesses of GAN-based oversampling for intrusion detection.**

***Index Terms*—Network Intrusion Detection, LSTM, Generative Adversarial Networks, Oversampling, Data Imbalance**

I. Introduction

As cyberattack techniques targeting network environments diversify and advance, traditional rule-based intrusion detection and prevention systems are becoming inadequate to address new threats such as Advanced Persistent Threats (APT). To overcome this limitation, research has been actively conducted on AI-based network intrusion detection systems based on machine learning. However, a significant issue in network traffic anomaly datasets is the severe imbalance in training data, where normal samples overwhelmingly outnumber attack samples. This study proposes an LSTM-based network intrusion detection model that utilizes GAN-based oversampling techniques to generate attack samples, and the performance of the proposed model is compared with existing machine learning models.

II. Background and Related Work

This section describes oversampling techniques to address the class imbalance problem and introduces two oversampling methods: SMOTE and GAN.

A. Oversampling

Oversampling is one of the representative methods for adjusting dataset proportions, alongside undersampling. It addresses data imbalance by increasing the sample size of the minority class. While oversampling improves recall by increasing the positive prediction ratio, it may reduce precision.

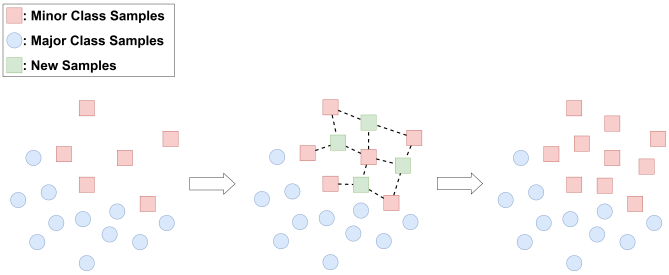

Fig. 1. SMOTE's oversampling process

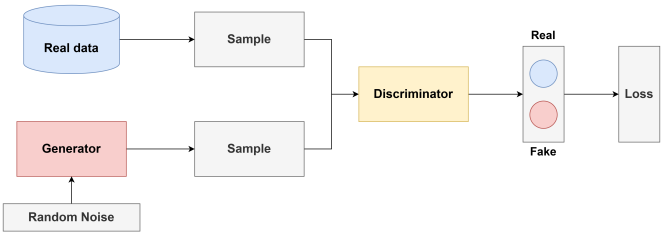

Fig. 2. GAN algorithm structure

B. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a classic oversampling technique that generates new data points by leveraging the k-nearest neighbors (KNN) of minority class data. Instead of merely replicating existing data, SMOTE creates new samples at a certain distance from existing points.

However, SMOTE can degrade classification performance in high-dimensional datasets. Network traffic datasets, typically comprising over 40 features, are high-dimensional. Thus, it is hypothesized that GAN-based oversampling may yield better results [1].

C. Generative Adversarial Networks (GAN)

GANs are composed of two deep neural networks: a generator, which creates synthetic data, and a discriminator, which distinguishes real data from synthetic data. Through their interaction, they generate data similar to actual data. As learning progresses in the generator model, the discriminator

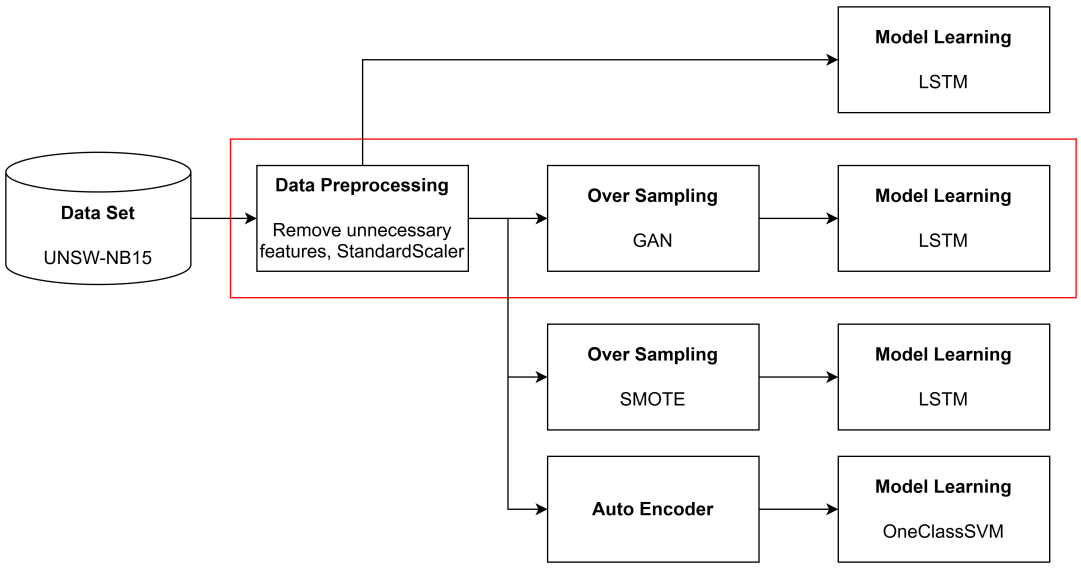

Fig. 3. Architecture of the proposed model and comparison group

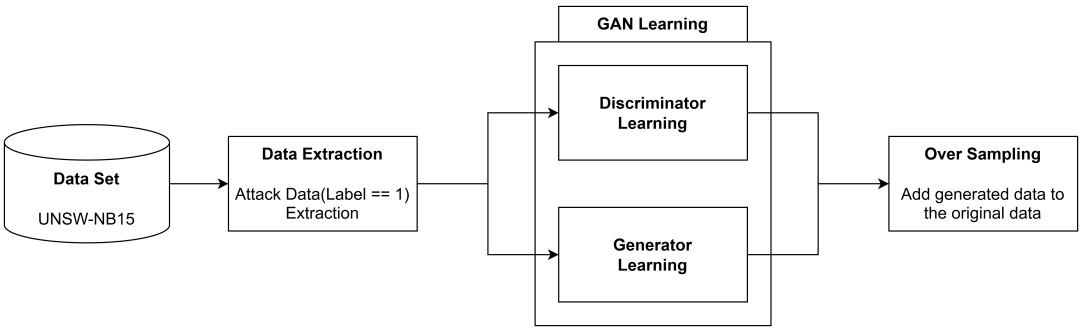

Fig. 4. Oversampling process using GAN

model becomes more robust to data transformation through the fake data that is continuously generated. Because of these characteristics, it is possible to learn to imitate data of any distribution through the GANs.

III. Proposed Method

The proposed system involves preprocessing the data, performing GAN-based oversampling, and training an LSTM model with the oversampled dataset. For comparison, three alternative models were also tested: 1) A simple LSTM model with preprocessing only; 2) An LSTM model using SMOTE for oversampling; 3) A model that addresses data imbalance with a One-Class SVM based on an autoencoder [2].

The experiments were conducted using the NSL-KDD, UNSW-NB15 dataset [3][4].

A. Data Preprocessing

Features such as transmission protocols (e.g., TCP, UDP) from the dataset were removed as they were deemed unsuitable for training. Standard scaler normalization was applied to ensure the mean of feature values was 0 and the standard deviation was 1.

B. GAN-Based Oversampling

The GAN-based oversampling process involves extracting attack data from the original dataset, training the GAN model to learn the distribution of attack data, and generating new synthetic samples. These newly generated samples are then combined with the original dataset for training.

C. LSTM Model Training

Given LSTM's capability to learn sequential and temporal patterns, it was employed for anomaly detection. The LSTM model aims to distinguish normal from abnormal patterns. Binary cross-entropy was used as the loss function, Adam as the optimizer, and a sigmoid activation function was applied.

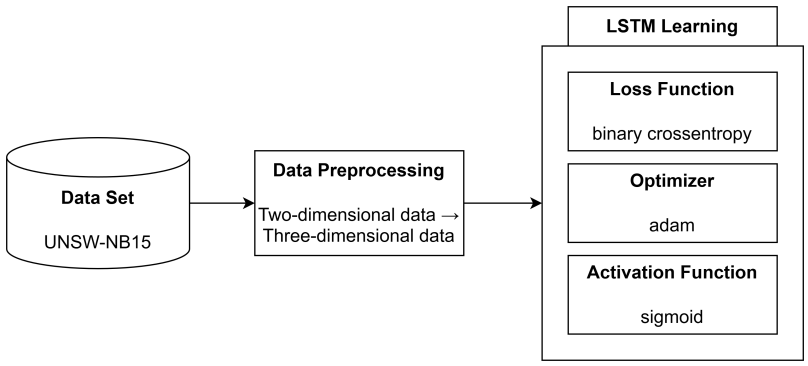

Fig. 5. LSTM learning process

IV. Evaluation

A. Evaluation Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The evaluation metrics for the model are Accuracy, Precision, Recall, and F1-Score, as defined by the equations above.

B. Evaluation Results

TABLE I
EVALUATION RESULTS AT UNSW-NB15

Model	Accuracy	Precision	Recall	F1-Score
LSTM (No Preprocessing)	0.8071	0.85	0.81	0.80
SMOTE + LSTM	**0.8447**	**0.86**	**0.84**	**0.84**
GAN + LSTM	0.4405	0.45	0.44	0.35
OneClassSVM	0.4876	0.59	0.49	0.40

Contrary to initial expectations, the simple LSTM model and the SMOTE-based LSTM model demonstrated the highest performance, with the latter slightly outperforming the former. The better performance of the SMOTE model indicates that resolving data imbalance through oversampling significantly enhances performance.

However, the GAN + LSTM model significantly underperformed, achieving F1-scores of only 0.35 and 0.38. This discrepancy highlights challenges in generating high-quality synthetic samples that align with the data distribution in the datasets. Another major factor contributing to the poor performance of the GAN-based model could be suboptimal hyperparameter settings or insufficient training of the GAN.

TABLE II
EVALUATION RESULTS AT NSL-KDD

Model	Accuracy	Precision	Recall	F1-Score
LSTM (No Preprocessing)	0.7817	0.80	0.80	0.78
SMOTE + LSTM	**0.7821**	**0.80**	**0.80**	**0.78**
GAN + LSTM	0.5177	0.39	0.46	0.38
OneClassSVM	0.5208	0.71	0.58	0.46

The One-Class SVM model also exhibited poor performance. While it effectively captures outliers in certain scenarios, it struggles with the high-dimensional dataset.

V. Conclusion

As mentioned in the results, in the case of current models, it is believed that meaningful results can be obtained when hyperparameter tuning is more appropriate.

As a way to increase the performance of the GAN + LSTM model other than hyperparameter tuning, we propose a method of transforming the dataset. LSTM was used in this project, but non-time series data was simply extended to three dimensions to be used like time series data. Instead of this method, it is expected that better performance can be derived if the dataset is transformed in a way that can utilize LSTM, such as a method of separating each transmission/reception data and forming two layers.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00217689) and the MSIT(Ministry

of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023-00259967) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] Rok Blagus et al, “SMOTE for high-dimensional class-imbalanced data,”Blagus and Lusa BMC Bioinformatics 2013, 14:106

[2] Byeoungjun Min et al, “Network Intrusion Detection with One Class Anomaly Detection Model based on Auto Encoder,” Journal of Internet Computing and Services(JICS) 2021. Feb.: 22(1): 13-22

[3] M. Hassan Zaib, ”NSL-KDD,” Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/hassan06/nslkdd

[4] D. Wells, ”UNSW-NB15 Dataset,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/mrwellsdavid/unsw-nb15