Privacy-Preserving Generative Adversarial Network-Based Data Synthesis for Intelligent Intrusion Detection Systems

Sun-Jin Lee
Department of Future Convergence Technology Engineering
Sungshin Women's University
Seoul, Korea
220237017@sungshin.ac.kr

Do-Eun Kim
Department of Convergence Security Engineering
Sungshin Women's University
Seoul, Korea
20221080@sungshin.ac.kr

Il-Gu Lee
Department of Future Convergence Technology Engineering
Sungshin Women's University
Seoul, Korea
iglee@sungshin.ac.kr

***Abstract*—With the development of the ICT industry, endpoint security issues are emerging. To quickly detect and respond to evolving attacks, it is necessary to improve threat detection performance without degrading endpoints' performance. Conventional data learning techniques have the problem of severe performance degradation when learning unbalanced and sparse datasets. In this paper, we propose a privacy-preserving generative adversarial network (PP-GAN) technique to solve the problems of conventional methods. In addition, we minimize damage caused by dataset leakage by filtering out privacy features that may cause information leakage in advance. According to the experimental results, even if the dataset is unbalanced and sparse, we improved the accuracy by 29% compared to the model that did not apply GAN by augmenting the dataset with privacy data deleted, achieving an average attack classification accuracy of 75.75%. In addition, we reduced data leakage by about 10% while reducing the accuracy by 0.55% compared to the general GAN-based dataset generation model.**

Keywords—Generative adversarial network, cyberattack, classification, privacy

I. INTRODUCTION

As all industries become digitalized, network complexity is increasing significantly. In particular, as the network environment changes to a dense network environment where multiple nodes participate in communication, cybersecurity issues in the data industry utilizing networks are increasing. In 2017, a cyberattack targeting a Renault factory caused the production system to be shut down, causing damage to the company worth millions of euros [1]. Even recently, attacks targeting Saint-Gobain, SolarWinds, and Merck have continued to occur, showing that the scope and impact of cyberattacks are gradually expanding.

A commonly used method for effectively detecting network intrusions is the network intrusion detection system (IDS). IDS is a system that analyzes network traffic to detect abnormal signs and discover malicious packets. Various attacks can be analyzed using IDS, and malicious patterns can be detected using a signature database. Afterward, suspicious or malicious network behavior can be defended, providing a beneficial and safe environment for individuals or businesses [2]. Existing machine learning-based IDSs have needed help to detect accurate attacks due to the diversity of network traffic and the lack of training datasets labeled with attack types [3]. In addition, as attack methods become more sophisticated and attack technologies develop, it is becoming increasingly difficult to detect and classify various types of attacks accurately.

Accordingly, research is being conducted to improve the performance of IDS by incorporating machine learning [4,5]. A large amount of data is required to learn a machine-learning model. If training is performed with a small amount of data, the imbalance of the data causes decision boundary deviation, which leads to a decrease in classification performance [6]. This phenomenon occurs more frequently in the case of traffic samples from unknown applications [7]. Research has been conducted to increase the amount of data by utilizing generative AI models to solve the data sparsity problem [8]. However, previous studies often did not consider the imbalance and sparsity of data during the data generation process and did not consider security when increasing the amount of data.

Therefore, in this paper, we propose a privacy-preserving generative adversarial network (PP-GAN). This balanced dataset generation framework considers security using the BoT-IoT dataset, a representative IoT traffic dataset. The proposed method solves the data imbalance problem while maintaining security and generating attack labels that need to be improved.

The main contributions of this paper are as follows.

- To improve the performance of the AI model, the GAN model was used to augment and balance network traffic data, improving the detection rate by up to 29%.
- Our approach strikes a balance between privacy protection and accuracy. By selectively excluding privacy information, we prevent personal information leakage by approximately 10%, with only a negligible average accuracy decrease of 0.55%
- By excluding and augmenting only privacy information, memory usage, and latency overhead are minimized.

The remainder of the paper is organized as follows. Section II analyzes an attack classification system based on data augmentation. In Section III, we propose PP-GAN to improve the performance of network attack traffic classification and enhance security. In Section IV, we analyze experiments and results, and in Section V, we conclude.

II. RELATED WORK

As network density has increased in recent years, the importance of cyber security through networks has increased. Real-time detection and analysis methods of network traffic are being utilized to prevent network intrusions. Network traffic analysis is essential for preventing congestion and detecting malicious packets through anomaly detection.

Table 1. Preliminary research on attack detection using GAN

Ref.	Techniques	Limitation	Data Balancing	Security Improvement
[9]	• Proposal of GAN-based model TDCGAN (Triple Discriminator Conditional GAN) • Improved minority class detection rate	• Consists of three discriminators, increasing computational cost and complexity • Only whether or not there is an attack is judged using binary classification	○	×
[10]	• Proposal of a generative deep learning model for IoT-23 dataset • Improved minority class detection rate • Prevention of zero-day attacks	• Binary classification does not cover various attack types • Not considering the information leakage issue.	○	×
[11]	• Proposal of GAN-based model BSDGAN (Balancing Sensor Data Generative Adversarial Networks) • Improving imbalance problems in the initial learning process using autoencoders	• Increased overhead due to structural complexity • Focuses only on a specific domain: human activity recognition sensors • Performance is evaluated only for binary classification.	○	×
[12]	• Proposal of an IDS combining semi-supervised learning and adversarial autoencoder	• Improve low accuracy • Only binary classification is provided	×	×
[13]	• Proposal of MalGAN algorithm to generate examples of hostile malware that bypass detection models based on black box machine learning • Reduce detection rate to close to 0	• Low response to new model attacks • It does not consider personal information leakage due to data generation	×	×
[14]	• Proposal of IDSGAN based on Wasserstein GAN to generate adversarial attack to bypass black box intrusion tower system • Generate adversarial attack samples with a high bypass rate	• Provide binary classification only • Failure to consider security	×	×
Our Model	• Augment and balance network traffic data using the GAN model to improve the performance of artificial intelligence models • Prevent personal information leakage by selectively excluding personal information		○	○

Table 1 summarizes previous papers that studied cyberattack detection and classification using GAN.

Jamoos et al. [9] proposed a GAN-based data balancing method using multivariate time series data generated by the system. They improved the detection rate of minority classes by learning the characteristics of intrusion samples of minority classes and generating balanced datasets. However, they measured binary attack classification, and the proposed generative model has the problem that it is computationally expensive and complex because it uses three discriminators.

Abdalgawad et al. [10] proposed generative deep learning methods such as AAE (Adversarial Autoencoders) and BiGAN (Bidirectional Generative Adversarial Networks) that can automatically detect and classify IoT cyberattacks. This method identified and classified attacks; the F1 score was 0.99 when detecting known attacks and 0.85 when detecting unknown attacks. However, this paper is specialized only for the IoT-23 dataset, and when other data is used, there is a problem that the bias of the dataset deteriorates the model performance.

Li et al. [11] proposed a new generative model, BSDGAN, which could significantly impact the field. Based on GAN, their model effectively solved the imbalance problem by generating a minority-class human activity sensor dataset. However, the model's complexity, particularly due to the combination of autoencoder, poses a potential challenge to its widespread adoption.

Kazuki et al. [12], a study that generates a dataset without considering the conventional balancing, used a semi-supervised learning algorithm AAE that integrates GAN with an autoencoder (AE) to build an intrusion detection system. Even though data augmentation was performed with only 0.1% of labeled data, binary classification showed low accuracy.

Hu et al. [13] proposed a GAN-based algorithm called MalGAN to generate adversarial malware examples that bypass black-box machine learning-based detection models. This method reduced the detection rate to almost 0 compared to the conventional gradient-based adversarial sample generation method. Still, it has the limitation of low adaptability when new attack methods appear.

Lin et al. [14] proposed a generative adversarial network IDSGAN framework that generates adversarial malicious traffic that bypasses Wasserstein GAN-based intrusion detection systems. However, although the traffic generated by

this method showed a high IDS bypass rate, security was not considered when generating malicious traffic.

Recently, various studies utilizing machine learning technology have been proposed. Accurate attack detection and classification are becoming important due to advanced attack methods and increasing types, and research is actively being conducted to generate attack datasets using machine learning and to balance datasets to improve detection performance. Among the many methods proposed to improve attack detection and classification performance, the most representative one is balancing the normal and attack datasets. Since the data of real industrial networks are usually unbalanced, the unbalanced dataset must be improved to improve performance [15]. This is because deep learning models can overfit many samples and underfit a small number of samples, leading to inaccurate network traffic detection and classification. However, previous studies have focused only on performance improvement from the network analysis perspective. Since attackers can use network analysis for malicious purposes, research that considers security is necessary.

III. Proposed Scheme

In this paper, we introduce a novel attack classification system based on a GAN model that considers privacy. This innovative approach not only enhances network traffic classification performance but also significantly improves security. The GAN-based dataset augmentation method, a key component of our system, can boost classification performance in environments with limited device storage capacity or uneven traffic by attack type.

The conventional GAN method does not consider privacy issues when augmenting the dataset and uses the dataset of all features as is. Therefore, if information related to personal characteristics is included in the dataset, there is a risk that an attacker can identify privacy features in the final dataset. The proposed model identifies and deletes privacy elements before applying GAN to the dataset and learns them to compensate for the personal information leakage problem. In addition, it can classify traffic attack types with high performance even with a small dataset.

The configuration diagram of the proposed model is shown in Fig. 1.

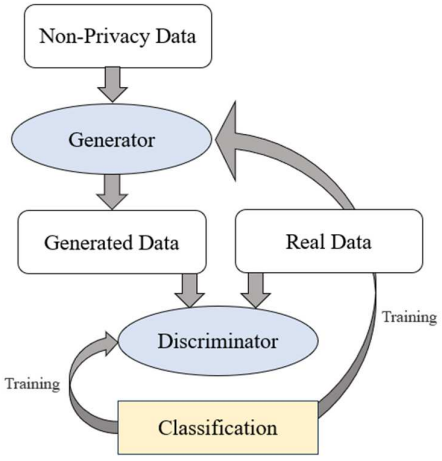

Fig 1. PP-GAN architecture

Unlike general GAN models, PP-GAN augments data only for data determined to be non-privacy. Features related to personal information can be manually selected by dataset providers and users. At this time, the Generator generates new data with a distribution similar to the existing dataset and augments the data by receiving a random noise vector. The Discriminator distinguishes whether the given data exists or is generated by the Generator. The competition between the Generator and the Discriminator improves the GAN model, and in particular, it generates a dataset close to reality when privacy features are deleted. When the proposed model is applied, the dataset distribution that is differentiated from the existing model is shown in Fig. 2.

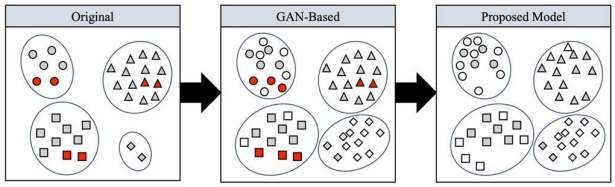

Fig 2. Example of dataset distribution of conventional and proposed models

In general network traffic datasets, there is bound to be an imbalance between normal and attack logs, and bias is bound to occur depending on the attack type. In addition, there are features (red blocks in Fig. 2) that can identify individuals in network traffic. In the real world, datasets collected have a few labels that make up most of the dataset, and some labels make up a tiny proportion, so the learning performance of the dataset is degraded. The GAN-based classification method balanced the dataset labels. Still, in the process of oversampling the dataset, unique privacy information on the labels is also generated, raising personal information issues. In contrast, the model proposed in this study minimizes privacy issues in the GAN model, balances the dataset, and ultimately creates a dataset that can enhance learning performance and security.

IV. Evaluation and Analysis

A. Evaluation environment

In this section, we evaluate the performance of the proposed GAN-based privacy-preserving network packet augmentation model. The dataset used in the experiment is the BoT-IoT dataset [16]. The BoT-IoT dataset is a dataset that contains data on normal and botnet traffic in a realistic network environment. The existing learning dataset and the oversampled proposed model's learning dataset for each label are shown in Table 2.

Table 2. Configuring label-specific learning datasets in BoT-IoT

Category	Original Dataset	GAN-based oversamling Dataset
Normal	370	100000
DDoS	1541315	100000
DoS	1320148	100000
Reconnaissance	72919	100000
Theft	65	100000

The attack classification labels of the dataset were multi-classified according to the categories the dataset had. The

security, accuracy, latency, and memory usage were measured in the same environment to compare with the existing non-privacy model. The experimental evaluation was performed in the Intel(R) Core™ i9-10850K CPU environment @ 3.60GHz, NVIDIA GeForce GT 1030 Graphics, 32.0GB RAM, and Python version 3.9.16.

We generated and evaluated datasets by increasing the balance rate by 0.1 from the original dataset 0 to the fully balanced dataset 1. The augmentation method divided the dataset into attack categories, augmented it considering the balance rate, and merged it to complete the augmented dataset. Then, we used the torch.nn module of the Pytorch library in Python to implement the model. ReLU was used as the hidden layer activation function of the generator and discriminator, and Tanh and Sigmoid were used as the output layer activation functions of the generator and discriminator, respectively. Optimization was performed using the Adam optimizer, and the learning rate was set to 0.0008. For the loss function, binary cross entropy was used to maximize the discriminator's prediction so that the data generated by the generator was discriminated as actual data, and the difference between the discriminator's prediction for the actual data and the prediction for the generated data was minimized.

B. Evaluation Results and Analysis

In this experiment, the performance of the proposed and existing models was compared using five evaluation indicators: security, accuracy, latency, efficiency, and memory usage. Learning was performed by increasing the balance ratio of the dataset from 0 to 1. By measuring the performance according to the data balance ratio, we aimed to compare the proposed model with the conventional model when a balanced dataset was created through data augmentation.

1) Security

The classification accuracy of the data that the user wants to protect was measured to evaluate the security. Among the BoT-IoT datasets, the feature that the user wanted to protect was set to 'state.' The state feature means the transaction state and consists of INT, RST, URP, REQ, and RIN. It is assumed that normal users classify the attack label, and attackers extract the state feature to identify the status of the user node and attack it. In the case of the conventional GAN model, as the label balance ratio increases, the performance of the general classification label gradually increases, and the classification performance of the protected feature is 100%. However, in the case of PP-GAN, the security was improved by reducing the accuracy of the protected feature by about 10% while minimizing the degradation of the performance of the general classification label.

2) Accuracy

Accuracy was calculated by dividing the number of correct data by the total number of data. The accuracy formula is as follows:

$$Accuracy(\%) = \frac{Correctly\ classified\ data}{Total\ data} \times 100 \qquad (1)$$

Fig. 3 shows the accuracy of the proposed and general GAN models as the data balance ratio increases. The logistic regression algorithm was used to measure the accuracy.

According to Fig. 3, as the balance ratio increases, the accuracy of both the proposed model and the conventional model increases, and on average, the conventional model shows an accuracy that is 0.51% higher than that of the proposed model.

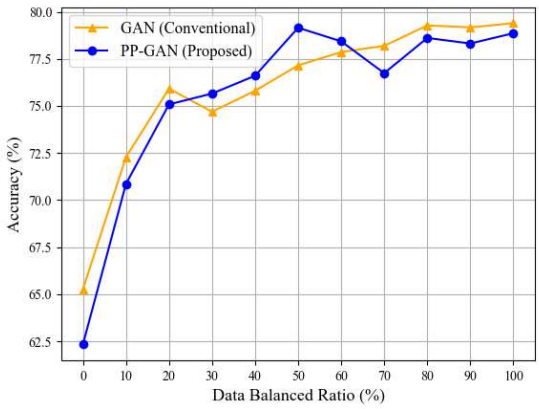

Fig 3. Accuracy per data balancing rate

3) Latency

Fig. 4 compares the delays in the data generation process of the general GAN model and the proposed model according to augmentation, considering the data balance ratio. As the balance ratio increases, the amount of data sets to be augmented increases, increasing the delays of both the conventional and proposed models. Unlike the conventional model, the proposed model is augmented, excluding privacy features, so the time is shortened by 1.81% compared to the conventional model in this process.

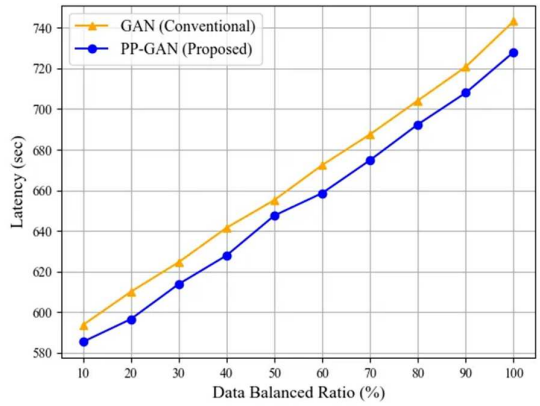

Fig 4. Latency per balancing rate

4) Memory Usage

Fig. 5 shows the conventional and proposed models' memory usage according to the data balance ratio. As shown in Fig. 6, as the balance ratio increases, the amount of data sets generated increases, increasing memory usage. In addition, the proposed and general models have an average difference of 7.29% and show almost the same increase trend.

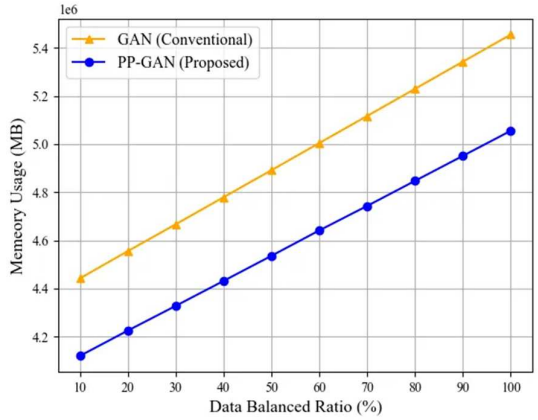

Fig 5. Memory Usage per balancing rate

5) Efficiency

The efficiency was measured to find the optimal point of the conventional and proposed models. The efficiency measurement method is shown in Equation (2). In the case of complexity, it was measured by memory usage rate.

$$Efficiency = \frac{Performance}{Complexity} = \frac{Accuracy}{Memeory\ Usage} \quad (2)$$

The efficiency of the conventional and proposed models is shown in Fig.6.

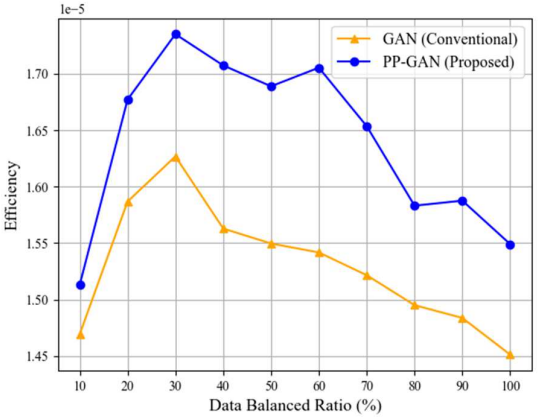

Fig 6. Efficiency per balancing rate

Both models demonstrate a trend where efficiency decreases as the balance ratio increases, a pattern attributed to the augmented dataset's increased memory usage. The conventional model shows high value of 1.6268 at 30% balance ratio, before gradually decreasing to its lowest value at 100%. The proposed model's efficiency increases with the balance ratio, reaching its peak of 1.7351 at 30% and gradually declining to its lowest point at 100%. This clear impact of the balance ratio on model performance enlightens you about the crucial role of dataset augmentation in influencing efficiency.

When comparing the increase in performance and complexity, both the proposed model and the existing model show the highest efficiency when the balance ratio is 30%. However, the efficiency of the proposed model with 30% augmentation is 6.66% higher than that of the existing model.

V. Conclusion

With the advancement of technology, networks are becoming increasingly dense, and various industries are using nodes to collect data. However, when collecting data in the IDS area, the ratio of normal logs is huge compared to the attack logs, making accurate learning difficult. Previous studies have used GAN to consider the balance of data but have yet to consider security simultaneously. Therefore, in this paper, we propose PP-GAN, a framework for creating a dataset that is safe from information leakage threats compared to general GANs. We simultaneously consider data balance and security and measure the most efficient augmentation ratio. As a result of evaluating the performance using the BoT-IoT dataset, the proposed model showed that the accuracy was reduced by only 0.55% compared to the existing GAN model while improving the delay and memory usage by 1.81% and 7.29%, respectively, and showed the best efficiency at a balanced ratio of 40%, increasing the efficiency compared to the existing model. In this study, we assumed the user selects the privacy feature in advance. Still, in future studies, we plan to study a feature set that can minimize information leakage by defining an automated metric.

Acknowledgment

This research is supported by the Ministry of Trade, Industry and Energy (MOTIE) under Training Industrial Security Specialist for High-Tech Industry (RS-2024-00415520) supervised by the Korea Institute for Advancement of Technology (KIAT), and the Ministry of Science and ICT (MSIT) of Development of Core Source Technology for Information Protection (No. RS-2024-00437252), and under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS-2022-00156310) supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP).

References

[1] A. Becue, I. Praça, and J. Gama, "Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities," *Artif. Intell. Rev.*, vol. 54, 2021, doi: 10.1007/s10462-020-09942-2.

[2] T. Thangaraj, S. Sridevi, C. D. Chelliah, T. Chung, and A. Khan, "Performance analysis of machine learning algorithms in intrusion detection system: A review," *Procedia Comput. Sci.*, vol. 171, pp. 1251–1260, 2020, doi: 10.1016/j.procs.2020.04.133.

[3] A. Aleesa, B. Bahaa, A. Zaidan, and S. Nan, "Review of intrusion detection systems based on deep learning techniques: Coherent taxonomy, challenges, motivations, recommendations, substantial analysis and future directions," *Neural Comput. Appl.*, vol. 32, 2020, doi: 10.1007/s00521-019-04557-3.

[4] M. Eskandari, Z. Janjua, M. Vecchio, and F. Antonelli, "Passban IDS: An intelligent anomaly-based intrusion detection system for IoT edge devices," *IEEE Internet Things J.*, pp. 1–1, 2020, doi: 10.1109/JIOT.2020.2970501.

[5] I. Idrissi, M. Azizi, and O. Moussaoui, "Accelerating the update of a DL-based IDS for IoT using deep transfer learning," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, pp. 1059–1067, 2021, doi: 10.11591/ijeecs.v23.i2.pp1059-1067.

[6] H. Ding, L. Chen, L. Dong, Z. Fu, and X. Cui, "Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection," *Future Gener. Comput. Syst.*, vol. 131, pp. 240–254, 2022, doi: 10.1016/j.future.2022.01.026.

[7] P. Wang, S. Li, F. Ye, Z. Wang, and M. Zhang, "PacketCGAN: Exploratory study of class imbalance for encrypted traffic classification using CGAN," in *2020 IEEE International Conference on Communications (ICC)*, Dublin, Ireland, 2020, pp. 1–7, doi: 10.1109/ICC40277.2020.9148946.

[8] I. Sarker, M. Furhad, and R. Nowrozy, "AI-driven cybersecurity: An overview, security intelligence modeling and research directions," *SN Comput. Sci.*, vol. 2, 2021, doi: 10.1007/s42979-021-00557-0.

[9] M. Jamoos, A. M. Mora, M. AlKhanafseh, and O. Surakhi, "A new data-balancing approach based on generative adversarial network for network intrusion detection system," *Electronics*, vol. 12, no. 13, p. 2851, 2023.

[10] N. Abdalgawad, A. Sajun, Y. Kaddoura, I. A. Zualkernan, and F. Aloul, "Generative deep learning to detect cyberattacks for the IoT-23 dataset," *IEEE Access*, vol. 10, pp. 6430–6441, 2022, doi: 10.1109/ACCESS.2021.3140015.

[11] D. Li, D. Kotani, and Y. Okabe, "Improving attack detection performance in NIDS using GAN," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, Madrid, Spain, 2020, pp. 817–825, doi: 10.1109/COMPSAC48688.2020.0-162.

[12] K. Hara and K. Shiomoto, "Intrusion detection system using semi-supervised learning with adversarial auto-encoder," in *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*, Budapest, Hungary, 2020, pp. 1–8, doi: 10.1109/NOMS47738.2020.9110343.

[13] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on GAN," *ArXiv*, abs/1702.05983, 2017.

[14] Z. Lin, Y. Shi, and Z. Xue, "IDSGAN: Generative adversarial networks for attack generation against intrusion detection," in *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part III*, Springer-Verlag, 2022, pp. 79–91, doi: 10.1007/978-3-031-05981-0_7.

[15] N. Tran, H. Chen, J. Jiang, J. Bhuyan, and J. Ding, "Effect of class imbalance on the performance of machine learning-based network intrusion detection," *Int. J. Perform. Eng.*, vol. 17, no. 9, p. 741, 2021.

[16] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Gener. Comput. Syst.*, vol. 100, pp. 779–796, 2019.