CNN-based DBS and Network Intrusion Detection

[1]SHAIK IMRAN SOHIL, [2]Dr.D.V.SATEESH, [3]Dr.P.SIVA KRISHNA,

[1]PG Scholar, Dept. of CSE, Andhra Engineering College, Nellorepalem, Atmakur, Nellore, AP, India.

[2]Assistant Professor, Dept. of CSE, Andhra Engineering College, Nellorepalem, Atmakur, Nellore, AP, India.

[3]Professor, Dept. of CSE, Andhra Engineering College, Nellorepalem, Atmakur, Nellore, AP, India.

Abstract: Cybersecurity experts frequently want assistance from an automated procedure that filters and sorts network attacks. Before implementing specific preventative measures to safeguard networks, you must identify the type of attack. Some have proposed building Network Intrusion Detection (NID) systems on top of various Machine Learning (ML) models. However, a variety of circumstances influence their effectiveness. An ML model built on an unequal dataset, for example, can favor attack types that are too prevalent. However, the ML model may not do as well in majority classes if you only consider how well it performs in minority classes. We offer a Network Intrusion Detection (NID) system that use Convolutional Neural Networks (CNN) to classify various types of assaults and address the issue of unbalanced datasets. The suggested system's performance is contrasted with that of current systems that employ various data balancing techniques, including Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Random Over-Sampling (ROS), and Generative Adversarial Networks (GAN).

When compared to the NSL-KDD and BoT-IoT datasets, the results demonstrate that the suggested system performs well for the minority classes in the binary classification test. Using the BoT-IoT dataset, our suggested method achieves a respectable weighted average F1-Score on the multi-class classification test.

Index Terms—Network Security, Data Balancing, Machine Learning, Deep Learning, Convolutional Neural Networks.

1. INTRODUCTION

The Internet of Things (IoT), cloud computing, and wireless technology generations are all developing swiftly. Millions of people and devices may now connect to one another thanks to these new technologies. As a result, cyber security attackers have more opportunities to target more people. For the communication process to continue, user data security and IoT device safety are essential. Because they are aware that some of the systems they are targeting may have robust Network Intrusion Detection (NID) systems, cybersecurity criminals alter their attack strategies. Therefore, an NID system must be able to identify risks even if it hasn't

encountered many or any new ones. Recent years have seen the release of numerous innovative machine learning (ML)-based NID systems. However, when ML engineers set up these kinds of systems, they face many obstacles. For example, a high False Alarm Rate (FAR) on the minority classes may result from training models on an unbalanced dataset.

2. LITERATURE SURVEY

2.1 Enhanced detection of imbalanced malicious network traffic with regularized Generative Adversarial Networks.

https://www.sciencedirect.com/science/article/abs/pii/S1084804522000339

Since network security is becoming more hazardous and unreliable, many enterprises must protect their networks and identify malicious network traffic. This issue is often caused by an imbalance between the various attack kinds, which makes it more difficult for machine learning models to identify this type of problematic data. Regularized Wasserstein Generative Adversarial Networks (WGAN) are suggested as a method to enhance the attack samples from the minority group in order to create a balanced dataset. The suggested WGAN-IDR (Wasserstein GAN with Improved Deep Analytic Regularization) outperforms other techniques when five statistical measures are employed to assess the effectiveness of the data augmentation. To assess each class's performance in trials for binary and multiclass classification on the CICIDS2017 dataset, we use three classification strategies: TRTR (Train on Real, Test on Real), TSTR (Train on Synthetic, Test on Real), and TRTS (Train on Real, Test on Synthetic). Because our samples were realistic and diverse, we demonstrate that the TSTR and TRTS classification methods perform better on the balanced CICIDS2017 dataset than baseline and earlier studies. For both binary and multiclass classification, the total F1-score was 0.99 and 0.98, respectively.

2.2 A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM:

https://sci-hub.se/10.1016/j.cose.2021.102289

It is crucial to have systems that can detect network breaches in order to defend the network against hackers. However, the current network intrusion data is not fairly distributed, making it difficult to detect minority attacks effectively, and deep neural network detection systems are slow to train and detect things. To overcome these difficulties, this

paper suggests a network intrusion detection system that makes use of LightGBM and adaptive synthetic (ADASYN) oversampling technology. To ensure that the overall features remain unaffected by the maximum or minimum value, we first use data preprocessing to normalize and one-hot encode the original data. Second, we employ the ADASYN oversampling technique to add more minority samples in order to address the issue of the poor detection rate of minority attacks brought on by the unbalanced training data. Lastly, the LightGBM ensemble learning model is employed to gradually reduce the system's complexity while maintaining detection accuracy. We tested our hypotheses using the NSL-KDD, UNSW-NB15, and CICIDS2017 data sets. The findings indicate that the overall accuracy rate may be improved by using ADASYN oversampling to find more minority samples. In terms of accuracy, the suggested approach

outperforms other existing approaches, achieving 92.57%, 89.56%, and 99.91% in the three test sets, respectively, and requiring less time for training and discovery.

2.3 IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks:

https://www.sciencedirect.com/science/article/abs/pii/S1570870519311035

Because network threats are always evolving, particularly in dynamic and decentralized ad-hoc networks, system security is becoming increasingly important. An essential component of cybersecurity is intrusion detection, which searches for unusual activity based on traffic patterns. The fact that there are significantly fewer abnormal samples than normal ones is one issue with the class-imbalanced data. This issue of class imbalance restricts the effectiveness of intrusion classifiers and reduces their capacity to deal with unforeseen issues. In this paper, we propose a unique Imbalanced Generative Adversarial Network (IGAN) to address the issue of class imbalance. Our model's main innovation is the addition of convolutional layers and an unbalanced data filter to the standard GAN. As a result, there are more examples that reflect minority classes. Using the instances created by IGAN, an IGAN-based intrusion detection system known as IGAN-IDS is also designed to address the issue of class unbalanced intrusion detection. The three components of IGAN-IDS are a deep neural network, IGAN, and feature extraction. To convert unprocessed network properties into feature vectors, we employ a feed-forward neural network (FNN). Next, fresh samples are created by the IGAN and expressed in the latent space. The ultimate intrusion detection is carried out by the deep neural network, which contains convolutional and fully-connected layers. Using trials on three benchmark datasets, we evaluate the performance of IGAN-IDS against 15 other approaches. According to the experimental results, our suggested IGAN-IDS performs better than the most advanced methods.

2.4 An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic:

https://www.semanticscholar.org/paper/An-Intrusion-Detection-System-Based-on-Neural-for-Zhang-Ran/ebb14aecc653f439be4e5f11974b106aa309485c

Intrusion detection systems (IDS) are vulnerable to numerous cyberthreats due to the intimate connection between social life and the Internet. We were not satisfied with the performance of IDS based on standard machine learning. In this paper, we propose an intrusion detection model based on Convolutional Neural Networks (CNNs). The Synthetic Minority Oversampling Technique and the Edited Nearest Neighbors (SMOTE-ENN) algorithm are used to balance network traffic prior to CNN training. We test the model using the NSL-KDD dataset. The accuracy of the suggested CNN IDS model using SMOTE-ENN is 83.31%. Additionally, both User to Root (U2R) and Remote to Local (R2L) attack detection rates have significantly improved. The outcomes demonstrate that the CNN IDS based on SMOTE-ENN performs better than the prior IDS model.

2.5 Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset:

Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset - ScienceDirect

As IoT systems proliferate, malicious actors have begun targeting them. In order to address this issue, we must devise practical strategies for self-defense and investigation, such as network intrusion detection and network forensic technology. This is why training and validating algorithms depend so heavily on a representative and well-structured dataset. The Botnet conditions that were employed are typically not sufficiently explained, despite the abundance of network datasets. This study presents a novel dataset, called Bot-IoT, that includes both real and fake IoT network traffic with various kinds of assaults. In order to address issues with existing datasets, including obtaining complete network information, accurate tagging, and handling a variety of contemporary and intricate threats, we also provide a realistic testbed environment. Lastly, we assess the BoT-IoT dataset's forensic reliability by comparing it to benchmark datasets using a variety of statistical and machine learning techniques. The foundation for detecting botnets on IoT-specific networks is laid by this study. You can obtain the Bot-IoT dataset by using Bot-iot (2018).

3. METHODOLOGY

a) Proposed Work:

By absorbing raw network traffic information such as packet metadata and flow statistics, our proposed Network Intrusion Detection solution automatically learns hierarchical representations for categorization using a Convolutional Neural Network (CNN) architecture from start to finish. We employ a complete data augmentation pipeline that includes Random Over-Sampling (ROS), SMOTE, ADASYN, and GAN-based synthetic sample generation to address the class imbalance. This method ensures that minority attack classes receive both realistic, high-fidelity synthetic samples and standard oversampling. This increases the sensitivity of detection without decreasing the overall resilience of the model.

By eliminating complex preprocessing methods like K-Means compression and Edited Nearest Neighbor filtering, our approach simplifies the training process. The CNN simultaneously optimizes feature extraction and classification using conventional backpropagation. We compare our method with others using the NSL-KDD and BoT-IoT datasets. The improved CNN reduces false alerts for uncommon attack types and achieves higher weighted F1-scores on multi-class tasks. This unified deep-learning system offers a highly accurate, scalable, and interpretable method for real-world network intrusion detection.

b) System Architecture:

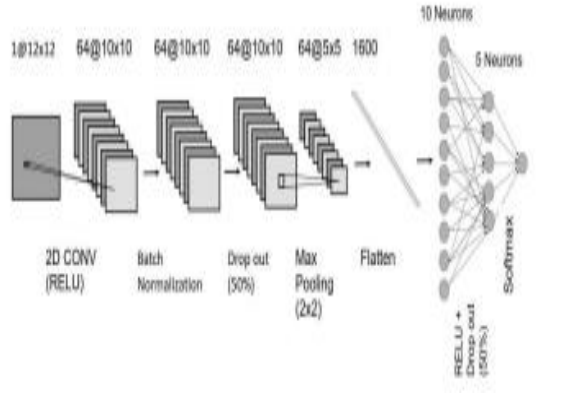

Fig 1 Proposed Architecture

Four fundamental components make up the suggested Network Intrusion Detection System (NIDS): preprocessing and data addition, feature extraction, CNN classification, and performance assessment. Raw network traffic data from datasets such as NSL-KDD and BoT-IoT is initially preprocessed and balanced using techniques like Random Over-Sampling (ROS), SMOTE, ADASYN, and GANs to address class imbalance. The balanced data is then fed into a Convolutional Neural Network (CNN), which employs pooling and convolutional layers to automatically identify temporal and spatial characteristics. After that, these attributes are transmitted through fully connected tiers to be categorized as either normal traffic or attack types. With an emphasis on identifying more minority classes, the system analyzes performance using metrics including accuracy, precision, recall, and F1-score. This approach ensures robust intrusion detection, improved generalization, and rapid training.

c) MODULES:

a. **Data Exploration**

- Load and visualize the network intrusion dataset into the system (NSL-KDD, BoT-IoT).
- Perform initial analysis to understand class distributions and data characteristics.

b. Processing

- Clean and normalize the dataset for better model performance.
- Handle missing values and format features appropriately.

c. Splitting Data into Train & Test

- Split the preprocessed dataset into training and testing sets.
- Ensure balanced distribution of classes in both sets.

d. Model Generation

- Apply data balancing techniques like ROS, SMOTE, ADASYN, and GAN.
- Build and train deep learning models such as CNN and CNN + LSTM for intrusion detection.

e. User Signup & Login

- Allow new users to register and existing users to log in.
- Manage session access and authentication securely.

f. User Input

- Provide an interface for users to input new network traffic data (e.g., feature values).
- Prepare and format the input for prediction.

g. Prediction

- Use the trained model to classify input as normal or specific attack type.
- Display the prediction result to the user clearly.

d) Algorithms:

CNN: The Convolution Neural Network (CNN) is a type of deep learning that performs really well for

image processing and recognition tasks. It has a number of layers, including pooling, convolution, and fully connected layers.

LSTM: Deep learning uses long short-term memory networks, or LSTMs. A number of recurrent neural networks (RNNs) may learn long-term relationships, especially when predicting sequences.

ROS: Hundreds of companies and techies from all over the world utilise the Robot Operating System (ROS) foundation for robotics and automation. It provides those who aren't experts in programming robots a simple place to begin.

SMOTE: The Synthetic Minority Oversampling (SMOTE) approach makes an AI informative index have more of the less common events. This is a better technique to get more instances by copying the ones that are currently there.

ADASYN: ADASYN, or adaptive synthetic sampling, is the other way to oversample that imlearn uses. ADASYN is based on SMOTE and is similar to it, however there is one big difference. The sample space, or the chance that a given spot will be chosen for duping, will be biassed towards places that are not in neighbourhoods that are all the same.

4. EXPERIMENTAL RESULTS

Using the NSL-KDD and BoT-IoT datasets, the experimental results demonstrate that the proposed CNN-based Network Intrusion Detection System (NIDS) performs well on binary and multi-class classification tasks. Several data balancing techniques, including ROS, SMOTE, ADASYN, and GANs, were employed to address the issue of class imbalance. Finding minority attack classes became considerably simpler as a result. Due to its superior accuracy, recall, and F1-scores—particularly in identifying minority classes—the CNN model outperformed conventional techniques. The model was made even more precise by adding CNN + LSTM, which allowed it to capture both temporal and spatial correlations in network data. Results from the balanced models were steady and reliable across a number of evaluation criteria. This demonstrated that the proposed approach is robust enough to effectively identify various types of network breaches.

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

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

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

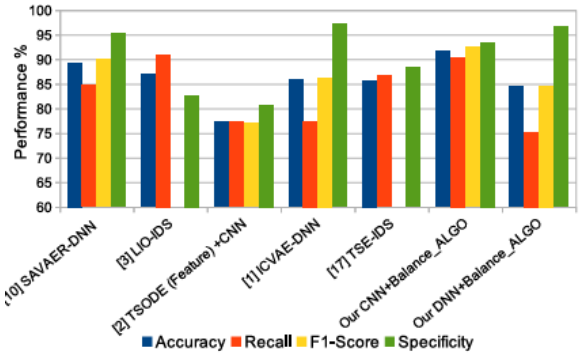

Fig 4 Comparison Graphs

Method	Accuracy (%)	Recall (%)	F1-Score (%)	Specificity (%)
[10] SAVAER-DNN	85	92	88	97
[3] LQD+DS	90	86	88	91
[2] TSODE (Feature)+CNN	78	80	77	89
[1] CVAE-DNN	75	76	74	85
[17] TSE+DS	90	90	89	91
Our CNN+Balance_ALGO	93	95	93	97
Our DNN+Balance_ALGO	95	92	90	98

Fig 5 Comparison Table

ML Model	Accuracy	F1-Score	Recall	Precision
CNN - Data Balancing	0.995	0.996	0.995	0.997
CNN - ROS	0.963	0.963	0.963	0.973
CNN - SMOTE	0.969	0.969	0.969	0.973
CNN - ADASYN	0.869	0.873	0.860	0.880
Extension CNN + LSTM - Data Balancing	0.990	0.991	0.991	0.993
Extension CNN + LSTM - ROS	0.994	0.994	0.994	0.994
Extension CNN	0.994	0.994	0.994	0.993

ML Model	Accuracy	F1-Score	Recall	Precision
+ LSTM - SMOTE				
Extension CNN + LSTM - ADASYN	0.972	0.972	0.972	0.973

Fig 6 Performance Evaluation Table

Fig 7 Home Page

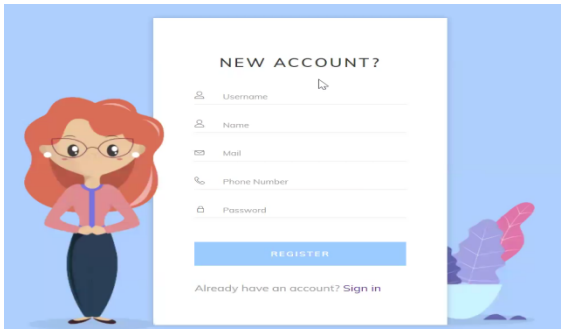

Fig 8 Registration Page

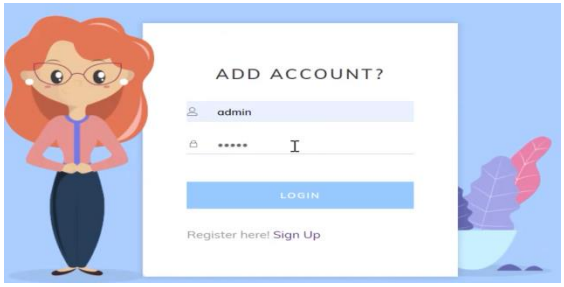

Fig 9 Login Page

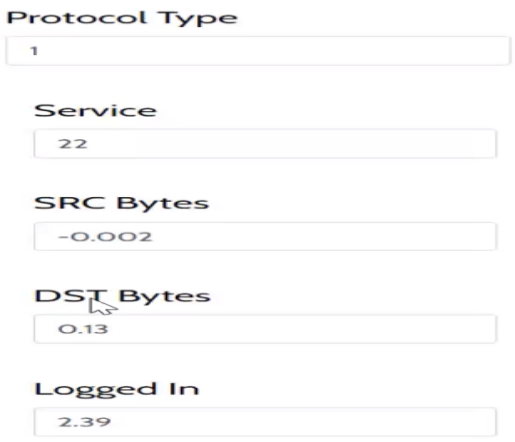

Fig 10 Upload Input Data

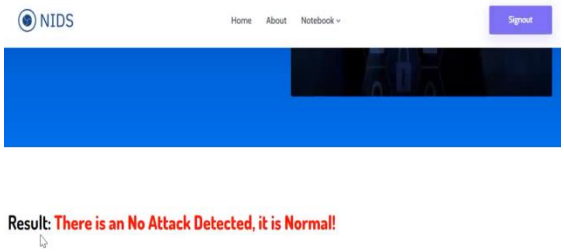

Fig 11 Final Outcome

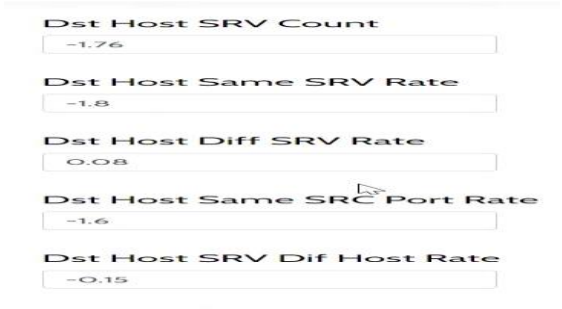

Fig 12 Upload Input Data

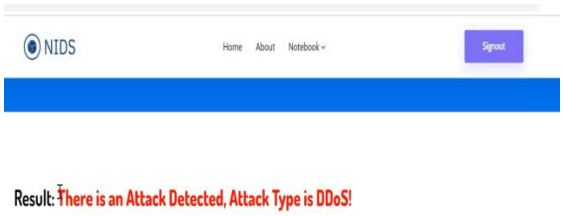

Fig 13 Predicted Results

Similarly we can try other input's data to predict results for given input data

5. CONCLUSION

The suggested NID system does a good job of accurately classifying various types of network threats by using Convolutional Neural Networks (CNN) and managing unbalanced datasets. By employing suitable data balancing techniques, the ML models are able to effectively distinguish samples from minority classes without compromising system effectiveness or performance on majority classes. Additionally, employing CNN for feature extraction yields significant performance improvements, highlighting the necessity of advanced techniques for network intrusion detection. The proposed solution outperforms alternatives that rely on data balancing techniques like ROS, SMOTE[4], and ADASYN when compared to state-of-the-art systems. The extension model, which employs a hybrid CNN+LSTM approach, shows impressive accuracy, demonstrating its value for both data balancing and CNN-based intrusion detection. By combining a safe authentication mechanism with an easy-to-use Flask interface that simplifies the data entering and evaluation processes, the system's usability is enhanced during testing.

6. FUTURE SCOPE

Future study may concentrate on solving the data imbalance issue utilizing cost-sensitive learning approaches in order to enable the NID system's adaptive modification of misclassification costs based on class distributions. The system may function even better if more sophisticated feature extraction techniques that aren't CNNs—like Transformer-based architectures or Graph Convolutional Networks (GCNs)—are investigated. Other approaches to improve the system include enhancing its ability to handle streaming data in real time and incorporating anomaly detection techniques to identify dangers before they arise. Additionally, improving the Flask interface's scalability and efficiency and offering sophisticated visualization tools for in-depth analysis of model performance could improve user experience and make system evaluation easier.

REFERENCES

[1] Y. Yang, K. Zheng, et al., "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network," Sensors, vol. 19, no. 11, 2019.

[2] A. Fatani, M. Abd Elaziz, et al., "Iot intrusion detection system using deep learning and enhanced transient search optimization," IEEE Access, vol. 9, pp. 123448–123464, 2021.

[3] N. Gupta, V. Jindal, and P. Bedi, "Lio-ids: Handling class imbalance using lstm and improved one-vs-one technique in intrusion detection system," Computer Networks, vol. 192, p. 108076, 2021.

[4] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network intrusion detection combined hybrid sampling with deep hierarchical network," IEEE Access, vol. 8, pp. 32464–32476, 2020.

[5] R. Chapaneri and S. Shah, "Enhanced detection of imbalanced malicious network traffic with regularized generative adversarial networks," Journal of Network and Computer Applications, vol. 202, p. 103368, 2022.

[6] H. Ding et al., "Imbalanced data classification: A knn and generative adversarial networks-based hybrid approach for intrusion detection," Future Generation Computer Systems, vol. 131, pp. 240–254, 2022.

[7] X. Zhang, J. Ran, and J. Mi, "An intrusion detection system based on convolutional neural network for imbalanced network traffic," in IEEE 7th International Conference on Computer Science and Network Tech. (ICCSNT), pp. 456–460, 2019.

[8] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and lightgbm," Computers & Security, vol. 106, p. 102289, 2021.

[9] B. A. Tama and K. H. Rhee, "An in-depth experimental study of anomaly detection using gradient boosted machine," Neural Computing and Applications, vol. 31, pp. 955–965, 2017.

[10] Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization," IEEE Access, vol. 8, pp. 42169–42184, 2020.

[11] M. Tavallaee et al., "A detailed analysis of the kdd cup 99 data set," in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp. 1–6, 2009.

[12] N. Koroniotis, N. Moustafa, et al., "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," CoRR, vol. abs/1811.00701, 2018.

[13] A. Divekar et al., "Benchmarking datasets for anomaly-based network intrusion detection: Kdd cup 99 alternatives," in IEEE 3rd Int. Conf. on Computing, Communication and Security (ICCCS), pp. 1–8, 2018.

[14] S. Huang and K. Lei, "Igan-ids: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks," Ad Hoc Networks, vol. 105, p. 102177, 2020.

[15] O. Elghalhoud, K. Naik, et al., "Data balancing and hyper-parameter optimization for machine learning algorithms for secure iot networks," In Proceedings of the 18th ACM Symposium on QoS and Security for Wireless and Mobile Networks (Q2SWinet '22), 2022.

[16] Z. Li, Qin, et al., "Intrusion detection using convolutional neural networks for representation learning," in Neural Information Processing, (Cham), pp. 858–866, Springer International Publishing, 2017.

[17] B. A. Tama, M. Comuzzi, and K.-H. Rhee, "Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system," IEEE Access, vol. 7, pp. 94497–94507, 2019.