IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.12486

AN ENSEMBLED NETWORK INTRUSION DETECTION SYSTEM USING AUTOENCODER TO RESOLVE DATA IMBALANCE

Dr.S.Kanmani[1]

Ms. M.R.Chaithra[2] , Mr. Balaji.A[3] , Mr. Gokulraj.R[4] , Mr. Sri Chandra Mouli[5]

Professor, Information Technology, Puducherry Technological University, Puducherry, India.[1]
B. Tech Student, Information Technology, Puducherry Technological University, Puducherry, India.[2-5]

Abstract: A Network intrusion detection system (NIDS) is a security technology that monitors network traffic for suspicious activity and alerts administrators or security personnel when potential threats are detected. The primary goal of a NIDS is to identify and respond to malicious activities such as unauthorized access, data theft, and other cyber-attacks. NIDS can be implemented using a variety of techniques, including signature-based detection, anomaly detection, and machine learning. By analysing network traffic in real-time, NIDS can provide an effective defence against cyber threats and help organizations protect their valuable assets from unauthorized access and data breaches. This abstract provides an overview of NIDS and its importance in network security.

Keywords: Autoencoder, Machine Learning, Ensemble Model

I.INTRODUCTION

Network intrusion detection systems play a crucial role in ensuring the security of modern computer networks, detecting and responding to a variety of cyber threats in real-time. However, one of the biggest challenges facing intrusion detection is dealing with imbalanced datasets, where the number of positive samples (i.e., attacks) is significantly lower than the number of negative samples (i.e., normal traffic). This can lead to a bias in the classification process, reducing the accuracy of intrusion detection systems.

To address this issue, we have explored the use of autoencoders, a type of artificial neural network, for synthetic data generation. By training an autoencoder on normal network traffic, it is possible to generate synthetic attack data that can be used to balance out the dataset and improve the performance of intrusion detection systems. However, autoencoders alone may not be enough to provide accurate predictions, as they may generate noise or fail to capture the full complexity of the network traffic.

To address these limitations, we have turned to ensemble models, which combine multiple classifiers to improve the accuracy and robustness of the intrusion detection system. In this paper, we propose an approach that combines autoencoders for synthetic data generation and an ensemble model for prediction. We demonstrate the effectiveness of this approach using real-world datasets, showing that our method outperforms traditional intrusion detection techniques in terms of accuracy, false positives, and false negatives. Overall, our results suggest that the combination of autoencoders and ensemble models holds great promise for the future of network intrusion detection.

II.MOTIVATION

Network intrusion detection system (NIDS) is a crucial component of any organization's cybersecurity infrastructure. Cybersecurity threats such as malware, viruses, and unauthorized access are constantly evolving, and organizations need to be vigilant in detecting and preventing these threats. NIDS helps in detecting and alerting the security team of any unusual network activity and intrusions, allowing them to take necessary actions to prevent any damage to the organization's network and resources. By deploying a NIDS, organizations can improve their network security posture

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.12486

and reduce the risk of data breaches, financial losses, and reputational damage. Additionally, NIDS can help organizations comply with regulatory requirements and standards related to cybersecurity. Developing and implementing an effective NIDS requires knowledge of network protocols, security threats, and technologies used to detect and prevent network intrusions. By exploring and developing a NIDS, you can gain valuable knowledge and experience in the field of network security and contribute to securing the digital assets of your organization or clients.

III.LITERATURE SURVEY

An Enhanced AI-Based Network Intrusion Detection System Using Generative Adversarial Networks [1], proposes a new approach to building an intrusion detection system using Generative Adversarial Networks (GANs). The system consists of an anomaly detection module based on a GAN and a classification module based on a Support Vector Machine (SVM). The GAN-based anomaly detection module learns the normal behaviour of the network traffic and identifies any deviations from it, while the SVM-based classification module identifies the type of intrusion based on the anomalies detected. The system outperforms existing intrusion detection systems on the NSL-KDD dataset in terms of detection accuracy and false-positive rate, highlighting the potential benefits of using GANs for anomaly detection in intrusion detection systems.

Tier-Based Optimization for Synthesized Network Intrusion Detection System [2], proposes a tier-based optimization approach for a synthesized network intrusion detection system (NIDS). The approach involves optimizing the different tiers of the NIDS separately to improve overall performance. The authors use a genetic algorithm to optimize the feature selection and parameter tuning of each tier, resulting in an optimized NIDS that provides better detection accuracy and fewer false positives. Experimental results on the NSL-KDD dataset show that the proposed approach outperforms existing NIDS in terms of detection accuracy, false-positive rate, and computational efficiency. The paper demonstrates the potential of tier-based optimization for improving the performance of NIDS and highlights the importance of optimizing the different components of the system separately to achieve better overall performance.

Fast anomaly identification based on multi-aspect data streams for intelligent intrusion detection toward secure industry 4.0 [3], proposes a fast anomaly identification approach for intelligent intrusion detection in Industry 4.0. The approach is based on analysing multi-aspect data streams generated by the network traffic, including packet size, inter-arrival time, and payload content. The authors use a deep learning model to process the data streams and identify anomalies in real-time. The proposed approach achieves high detection accuracy and low false-positive rate, making it suitable for real-world industrial applications. The paper highlights the importance of developing intelligent intrusion detection systems for Industry 4.0 and demonstrates the potential of using deep learning techniques for analysing multi-aspect data streams for anomaly identification.

IV.LIMITATIONS IN EXISTING SYSTEM

- **High False Positive Rate:**

One of the biggest challenges in IDS is to accurately distinguish between normal and abnormal behaviour. Many IDS generate a large number of false positives, which can overwhelm security administrators and decrease overall system efficiency.

- **Bias and Overfitting:**

Machine learning models can be biased towards certain types of data, leading to inaccurate results and false negatives. Overfitting can also occur, where the model is too closely tailored to specific data, reducing its ability to generalize to new data.

- **Dependence on Quality Data:**

AI-based IDS rely on large amounts of labelled data to train and validate the models. This data must be of high quality and representative of normal and abnormal behaviour to ensure accurate results. If the data is biased or limited in scope, this can affect the accuracy of the system.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.12486

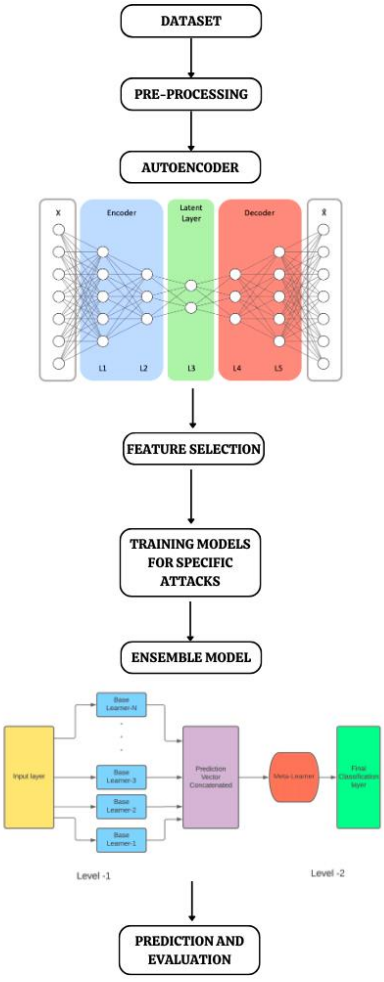
V.PROPOSED SYSTEM

The proposed system aims to detect various types of network attacks and anomalies by analysing network traffic data. The system utilizes autoencoder to generate synthetic data and machine learning algorithms such as Random Forest, Decision Tree, K-Nearest Neighbour (KNN), Support Vector Machine (SVM) for classification and prediction. The pre-processing step includes techniques such as removing duplicate data, handling missing values, and transforming categorical data into numerical data. The dataset is split into two parts, a training set and a testing set, using a 10% sampling ratio for the testing set. The pre-processing step then involves applying one-hot encoding to transform categorical data into numerical data, which is essential for machine learning algorithms to process the data.

Then the system applies various machine learning algorithms such as K-Nearest Neighbour (KNN), Decision Tree, and Random Forest to the NSL-KDD dataset to build models for network intrusion detection. The system trains the models using the training dataset and evaluates their performance on the testing dataset. The models are then stack ensembled for evaluation.

Overall, the proposed system aims to provide an accurate and reliable solution to the problem of network intrusion detection. The system's effectiveness is expected to enhance the security of networks, providing better protection against cyber threats.

VI.DESIGN DIAGRAM FOR PROPOSED SYSTEM



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.12486

VII.EXPERIMENT AND EVALUATIONS

Data set description:

NSL-KDD (NSL stands for Network Security Lab and KDD for Knowledge Discovery and Data Mining) is a dataset commonly used for network intrusion detection research. It is an updated version of the KDD Cup 1999 dataset, which was widely used in the past. NSL-KDD was released in 2009 and aims to address some of the shortcomings of the KDD Cup 1999 dataset, such as the lack of diversity in the attack types and the presence of redundant records. NSL-KDD includes both normal network traffic and various types of attacks, such as DoS, probing, user-to-root (U2R), and remote-to-local (R2L) attacks. The dataset contains 41 features that capture information about the network connections, such as the source and destination IP addresses, protocol type, service type, and duration of the connection. It consists of a training set of 125,973 records and a test set of 22,544 records. NSL-KDD has been widely used in research for evaluating and comparing the performance of various intrusion detection systems and techniques.

Implementation and Evaluation:

Implementation and Evaluation: In the experiment, accuracy, precision, recall, and F1-score were the four metrics we used to assess the performance of AI models. Accuracy is a term frequently used to evaluate the effectiveness of AI models. It represents the percentage of outputs that are accurately predicted. Precision shows the proportion of positive values reliably inferred by the model for a specific class in a data set, whereas Recall shows the percentage of data containing positive values properly inferred by the model. The harmonic mean of Precision and Recall is the F1 score. The following definitions describe the formulas for these measurements:

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

2) Precision = $\frac{TP}{TP + FP}$

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

4) F1-Measure = $2 \times \frac{Precision \times Recall}{Precision + Recall}$

Table 1

Cross Validation Results For The Test Data Set In NSL-KDD

Algorithm	*Attacks*							
	DOS				PROBE			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Random Forest	99.69%	99.66%	99.67%	99.69%	99.35%	99.27%	99.08%	98.98%
K-Neighbour	99.71%	99.67%	99.66%	99.67%	99.07%	98.60%	98.50%	98.55%
SVM	99.37%	99.10%	99.45%	99.27%	98.45%	96.90%	98.36%	97.61%
Ensemble model	99.80%	99.85%	99.69%	99.77%	99.25%	98.78%	98.94%	98.84%

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.12486

Algorithm	*Attacks*							
	R2L				U2R			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Random Forest	96.74%	95.32%	95.48%	97.02%	99.66%	98.34%	84.37%	87.73%
K-Neighbour	96.74%	95.32%	95.48%	95.40%	99.07%	93.14%	85.07%	87.83%
SVM	96.79%	94.85%	96.26%	95.52%	99.63%	91.05%	82.90%	84.86%
Ensemble model	97.25%	95.87%	96.33%	96.04%	99.74%	94.10%	87.34%	89.15%

VIII.CONCLUSION

In conclusion, the Network Intrusion Detection System proposed, uses AutoEncoder for Synthetic Data Generation and an Ensemble Model for Prediction has shown promising results in addressing the data imbalance issue and improving the classification performance of the system. The proposed approach generates synthetic data using an AutoEncoder and trains the Ensemble Model on a combination of real and synthetic data, leading to improved classification performance. The system also demonstrates the effectiveness of the Ensemble Model in combining the strengths of multiple classifiers to improve overall performance. This approach presents a promising solution to Network Intrusion Detection and could have significant practical implications in various applications.

IX.REFERENCES

[1] C. Park, J. Lee, Y. Kim, J. -G. Park, H. Kim and D. Hong, "An Enhanced AI-Based Network Intrusion Detection System Using Generative Adversarial Networks," in IEEE Internet of Things Journal, vol. 10, no. 3, pp. 2330-2345, 1 Feb.1, 2023, doi: 10.1109/JIOT.2022.3211346.

[2] M. A. Siddiqi and W. Pak, "Tier-Based Optimization for Synthesized Network Intrusion Detection System," in IEEE Access, vol. 10, pp. 108530-108544, 2022, doi: 10.1109/ACCESS.2022.3213937.

[3] L. Qi, Y. Yang, X. Zhou, W. Rafique, and J. Ma, "Fast anomaly identification based on multi-aspect data streams for intelligent intrusion detection toward secure industry 4.0," IEEE Trans. Ind. Informat., vol. 18, no.9, pp. 6503–6511, Sep. 2022

[4] B.Yan and G. Han, "Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system," IEEE Access, vol. 6, pp. 41238–41248, 2018.

[5] N.Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no. 1, pp. 41–50, Feb. 2018.

[6] C.Ieracitano, A. Adeel, F. C. Morabito, and A. Hussain, "A novel statistical analysis and autoencoder driven intelligent intrusion detection approach," Neurocomputing, vol. 387, pp. 51–62, Apr. 2020.

[7] J.Y. Kim, S. J. Bu, and S. B. Cho, "Malware detection using deep transferred generative adversarial networks," in Proc. Int. Conf. Neural Inf. Process., 2017, pp. 556–564.