L-014

Integration Of Machine Learning Techniques in a Hybrid Method for Network Traffic Anomaly Detection

Busireddy Mokshitha Reddy‡ Bhed Bahadur Bista‡

Eiichiro Kodama‡ Jiahong Wang‡

1. Introduction

In this technological world, as technology in every aspect is increasing day by day, mainly in online-based applications and services, in the same way, Cyberattacks are also increasing along with the technology. So, it has become prominent to detect the anomalies that are caused by the attacks in the traffic; this is crucial for the protection and coherence of the system, as the risks caused by the attacks become complicated. In this modern era, we know that machine learning algorithms have become powerful and beneficial tools for finding abnormal behaviour in network traffic.

We have proposed a hybrid method that unifies XG-Boost and the Auto-Encoders to tackle this. Relaying on the reconstruction errors, the abnormal behaviour can be identified by the Auto-Encoders and then sent into the XG-Boost. This gradient boosting model is for more accurate classification as it is a supervised learning model, which is more potent in feature extraction, and detection accuracy also increases. The dataset that is used is a combination of the four datasets, including CIC-IDS2017, CIC-IDS2018, CIC-IDS2019, and CIC-DOS 2017.

2. Related Work

The utilisation of deep learning methods has increased the popularity of anomaly detection in networks. Sreenivasa Rao et al. [1] proposed a Convolutional Neural Network (CNN) and GAN hybrid model. The model has been employed using the NSL-KDD dataset to find the anomalies in the traffic. This model is lacking when new threats occur. This CNN and GAN hybrid model cannot perform on the large data that is imbalanced. This Model is more complex. When this model is experimented with CIC-IDS2017, it can only detect outliers with 79% accuracy and 78% precision. In Sakurada et al. [2], a new strategy using Autoencoders was proposed. This model is employed in real-time IDS. This model flags outliers based on the errors given by the model. This model could reconstruct normal patterns and detect anomalies by high reconstruction error, but we only have binary labels (normal or anomaly), and there is no interoperability with a supervised classification technique for attack classification. Chen et al. [3] proposed an XGBoost model that has been employed in intrusion detection tasks in machine learning techniques. This classification is based on engineered features. It has robust performance compared to standard classifiers like SVM and Decision Trees, and Random Forest in terms of accuracy and speed, particularly for the structured datasets, and it does not natively detect anomalies without being labelled in advance. It does not perform well on an unknown zero-day attack, and it gives a weak generalization without any extraction.

‡ Graduate School of Software and Information Science, Iwate Prefectural University

3. Proposed Model

The Hybrid Auto-Encoders and XG-Boost approach is for the identification of anomalies when there are data breaches. This Hybrid method leverages the powerful classifying capacity of XG-Boost and Auto-Encoders with the vast feature extraction ability, and by taking reconstruction faults as a crucial categorisation characteristic, the methodology is intended to identify network traffic irregularities more precisely. The proposed hybrid method architecture is shown in Figure 1.

The following phases describe how the suggested approach operates:

3.1 Dataset accumulation

In this paper, we used CIC-IDS2017, CIC-IDS2018, CIC-IDS2019, and CIC-DOS 2017, and these are combined into one big dataset. The dataset is employed.

3.2 Cleaning up the Data

The data is pre-processed by managing data that is lacking & omitting superfluous characteristics. The purpose of this is to make sure that every characteristic is on the same scale.

3.3 Training of the Encoders

An automated encoder is trained using regular traffic data from networks, and it is employed for training to decrease reconstruction error for typical circumstances. An Auto-Encoder tends to compress so that it can rebuild the source data afterwards.

3.4 Determining the Reconstruction Errors

After training, the Auto-Encoder can recreate both normal and anomalous scenarios. For each sample, the reconstructed error is computed. High reconstruction mistakes are suggestive of possible anomalies.

3.5 Feature Extraction

Reconstruction errors are added to the original traffic characteristics to generate a new feature set that better highlights aberrant patterns.

3.6 XGBoost-Classifier

The reconstruction mistakes are sent into an XG-Boost classifier, which is trained to distinguish between normal and

abnormal traffic. XG-Boost's gradient boosting architecture improves prediction accuracy and handles data by focusing on hard-to-classify instances.

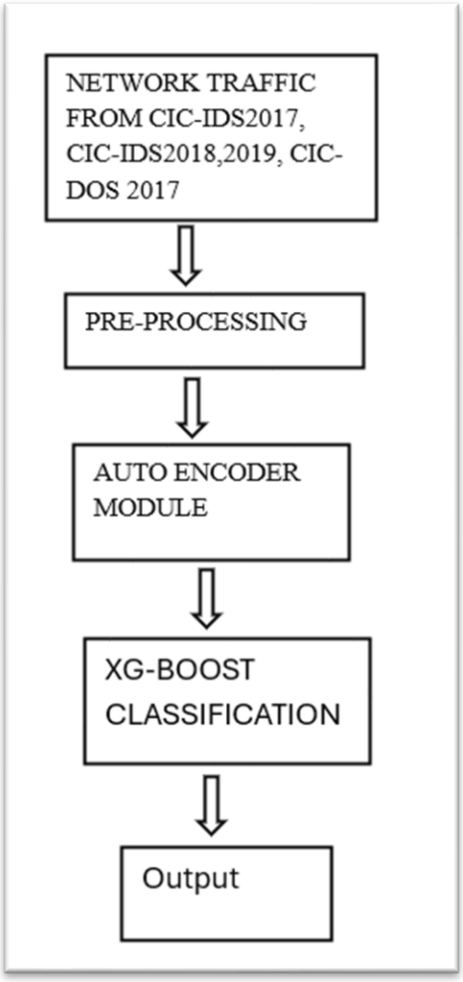

Figure 1. Proposed hybrid method Architecture

4. Results and Discussions

The developed hybrid anomaly detection model, combining Autoencoder (AE) and XG-Boost for feature extraction and classification, proved to be very effective in detecting the anomalies generated by different network attacks. The performance of the model was assessed based on standard classification measures: Accuracy, Precision, Recall, and F1-score. The experimental results of the related work and the proposed method are presented in Table 1.

The accuracy performance of AE–XGBoost was 98.86%, suggesting that the model accurately classified most of the normal and anomalous traffic instances. This Hybrid method achieved an accuracy of 98.47%, indicating a low rate of false positives, which is crucial in IDS systems, and this model can correctly identify an overwhelming majority of real attack activities with a recall of 98.42%, ensuring that little malicious activity can be missed.

Table 1 Performance evaluation of the previous work and the proposed work.

Methods	Accuracy	Precision	Recall	F1-Score
CNN-GAN model	79	78	79	76
Proposed AE-XG Boost	98.86	98.47	98.42	98.42

5. Conclusion

In conclusion, Auto-Encoders and XG–Boost hybrid method is effective and the ideal solution for anomaly detection when breaches take place in the network for modern attacks and for large, complex data. This strategy can be a solid framework for continuous development in the evolving cybersecurity landscape. Employing a hybrid approach in organisations can further improve security measures through the integration of streams of real-time data and the establishment of robust defences against zero-day attacks. It is essential to focus on the installation and integration of these advanced technologies within existing operational networks to ensure their efficacy and scalability. This research gives a path to address the complex challenges posed by cyberattacks in today’s digital environment.

References

[1] Vuda Sreenivasa Rao, R. Balakrishna, Yousef A. Baker El-Ebiary, Puneet Thapar, K. Aanandha Saravanan, and Sanjiv Rao Godla. "AI Driven Anomaly Detection in Network Traffic Using Hybrid CNN-GAN", Journal of Advances in Information Technology, Vol. 15, No. 7,2024.
[2] Sakurada, M., & Yairi, T., " Anomaly detection using autoencoders with nonlinear dimensionality reduction", Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, 4–11, 2014.
[3] Chen, T., & Guestrin, C., " XGBoost: A scalable tree boosting system", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, 2016.