Turkish Journal of Computer and Mathematics Education *Vol.10 No.02 (2019), 1063-1067*
DOI: https://doi.org/10.17762/turcomat.v10i2.13626

Research Article

Anomaly Detection in Network Traffic using Machine Learning and Deep Learning Techniques

Samir Rana
Dept of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002

Abstract
Due to the rise of sophisticated cyberattacks, network security has become an increasingly important field. One of the most common threats to the security of networks is network anomalies, which can cause system malfunctions and prevent them from working properly. Detecting such anomalies is very important to ensure the continued operation of the network. Deep learning and machine learning algorithms have demonstrated their ability to detect network anomalies, but their effectiveness is still not widely known. This paper presents an evaluation of the performance of three algorithms against the KDD-NSL dataset. This study aims to provide a comprehensive analysis of the various techniques used in deep learning and machine learning to detect network anomalies. It will also help improve the security of networks. The paper presents an evaluation of the performance of three algorithms against the KDD-NSL dataset. The three algorithms are the Support Vector Machine, the Random Forest, and the Artificial Neural Network. They will be compared with their accuracy, recall, and F1-score. The study also explores the impact of the algorithm's feature selection on its performance. The findings of the investigation will be used to inform the development of new techniques that can be utilized to enhance the security of networks. The KDD NSL dataset provides an ideal opportunity to analyze the performance of various algorithms for detecting network anomalies.
Keywords: Anomaly Detection, Network Anomaly, Machine learning, Deep Learning, SVM, ANN, Random Forest

Introduction

The rise of the internet and the increasing number of connected devices have made network traffic an essential part of our daily lives. We rely on it to carry out various activities, such as streaming videos and making payments online. However, as the amount of data that's transmitted over the network keeps growing, it has become harder to maintain its reliability and security. Due to the increasing number of attacks on the network, many companies and organizations are now becoming more concerned about the security of their networks.[1]

An important technique for protecting networks is anomaly detection, which is a process that can identify anomalous behavior in a wide range of network data. This type of detection can be applied to various aspects of network data such as user activity logs and network traffic. There are two types of techniques used for this type of detection: machine learning-based and statistical-based.[2]

Methods that are statistical-based can identify deviations from the expected behavior. On the other hand, methods that are machine learning-based use algorithms to learn from the patterns of normal behavior. A failure to detect network anomalies can lead to various issues, such as data breaches and system downtime. These can be caused by various factors, such as hardware failures, malicious attacks, and software bugs. If left undetected, these can affect an organization's operations and financial security.[3], [4]

A successful cyberattack can have severe consequences, such as the theft of sensitive data or the disruption of vital infrastructure. Apart from financial losses, a company's reputation can also be damaged. An organization's ability to detect network anomalies is very important to prevent cyberattacks and ensure the integrity of its networks. This process can be used to identify potential threats before they cause damage. In real-time, it can help prevent further issues by blocking suspicious traffic and isolating the affected systems.

The goal of this study is to analyze the performance of three different algorithms: the Support Vector Machine, the Random Forest, and the Artificial Neural Network. We will compare these with respect to accuracy, recall, F1-score, and precision. In addition, we will explore the impact of the feature selection on these algorithms' performance. This study aims to analyze the various techniques used for detecting network anomalies and their effectiveness in improving the security of networks.[5], [6]

The KDD NSL dataset, which is a widely used research tool in network security, contains traffic data that has been simulated to attack different types of DoS attacks. This was created by NIST to support research related to intrusion detection systems. The KDD NSL data set contains about 4.9 million records that have been labeled as either "normal" or "DoS, Probe, R2L, U2R". It features 41 features, such as service type, destination IP addresses, flags, and protocol type.

One of the main benefits of utilizing the KDD-NSL dataset is its vast number of labeled examples. This makes it an ideal training and testing ground for machine learning models. Furthermore, it contains a wide variety of

DOI: https://doi.org/10.17762/turcomat.v10i2.13626

attacks, which can be used to evaluate the algorithms' performance. The goal of this study is to analyze the performance of deep learning and machine learning algorithms when it comes to detecting DoS attacks. This type of attack is very common and can severely affect a company's operations. By focusing on this attack, we hope to gain a deeper understanding of how different techniques can identify it. The KDD-NSL dataset can provide us with an opportunity to thoroughly study deep learning algorithms and machine learning models when it comes to identifying network anomalies. The results of this research will be beneficial in helping organizations protect their networks from cyberattacks.

Literature Review

In order to identify anomalous traffic in a network, M. Mantere et al.[7] developed a method that combines decision trees and various features to classify the traffic. The proposed method can achieve an accuracy of 98.4%.
M. Naser et al.[8] proposed method combines the various statistical techniques used in network traffic analysis, such as clustering, anomaly detection, and principal component analysis. The evaluation of the proposed method revealed that it performed better than other methods when it came to detecting network anomalies.
To analyze the traffic features in a network, F. Iglesias et al.[9] used a university network's data to perform an analysis of the various features. They then used machine learning techniques to find the most accurate algorithm for detecting network anomalies.
To identify network anomalies, T. Andrysiak et al.[10] developed a method that uses the ARFIMA model, which is an autoregressive model for analyzing railway traffic data. Their method was evaluated against other techniques and revealed that it was more accurate at detecting anomalies.
M. Ding et al.[11] proposed a method that uses the PCA model to reduce the dimensionality of the data. It then uses the residuals of the model to detect anomalies. The proposed method was evaluated against other methods and revealed that it performed better than them.
Nie et al.[12] review the various applications of deep learning and machine learning in cybersecurity. They talk about the various tasks that these techniques perform in detecting threats, such as vulnerability analysis and intrusion detection. The paper also provides a comprehensive analysis of the issues that affect their development.
Naseer et al.[13] present a deep neural network-based method that can detect network anomalies. They then compare their approach with other methods by using the NSL-KDD data. The results of their evaluation revealed that their method is more accurate than the others.
Nguyen et al.[14] present a method that uses a PCA-based approach to detect network anomalies in an IoT network. The authors then use the collected data to analyze the anomalies and find them based on their residual errors. The evaluation of their method revealed that it can effectively identify anomalous traffic in the network.
Radford et al.[6] present a method that uses a recurrent neural network to detect network anomalies. They then compare their approach with other methods by using the NSL-KDD data. The results of their evaluation revealed that their method is more accurate than the others. The researchers evaluated the performance of the method against various network topologies and noise levels in the collected data.
Xin et al.[2] provides a comprehensive review of the various aspects of deep learning and machine learning techniques for cybersecurity. They discuss their applications in various areas, such as vulnerability analysis and intrusion detection. It also highlights their limitations and suggests future directions.

Anomaly Detection in network

An anomaly detection is a process that finds unusual patterns in the traffic in the network. It aims to identify potential security breaches or other issues that could affect the network. An anomaly detection solution provides various advantages to network administrators. It can help them identify potential security threats such as malware and hacking attempts. It can also help them identify system failures that could cause downtime. In addition, it can help them plan their resources and improve the performance of their networks.
Unfortunately, there are some limitations to network anomaly detection. One of the most challenging factors is distinguishing between abnormal and normal traffic. Since the behavior of the network can vary depending on various factors, such as the time of day and the user's behavior, it is not easy to define a standard for normalization.
Unfortunately, one of the biggest limitations of an anomaly detection solution is the high number of false positives. This can be caused by the mistake of identifying legitimate traffic as anomalous. It can be very time-consuming and costly to investigate. Also, it can't effectively detect attacks that are designed to evade detection.
An anomaly detection solution is useful for maintaining computer networks' security and performance. Although it has some limitations, deep learning and machine learning techniques can help improve its accuracy.

DOI: https://doi.org/10.17762/turcomat.v10i2.13626

Methodology

A. Data preprocessing

The quality of the data collected is a critical factor that affects the performance of deep learning and machine learning systems. In this study, the KDD NSL data will be preprocessed to enable the development of deep learning and machine learning algorithms. The three steps involved in the data preprocessing are data normalization, data cleansing, and data transformation.

i. Data Cleaning: The first step in the data preprocessing process is to remove the unnecessary and noisy data from the KDD NSL dataset. This will help improve the accuracy and reduce the complexity of the data. In addition, we will also remove duplicates and invalid data to ensure that the data is complete and consistent.

ii. Data Transformation: The data transformation step is the next step in the preprocessing process. In this process, the categorical variables will be transformed into numerical ones, which will allow the deep learning algorithms to perform their operations. One-hot encoding will also be used to analyze the relationships between the various variables.

iii. Data Normalization: In the final step of the data preprocessing, data normalization is performed to ensure that the data is on a similar scale. This process is carried out using the Z-score normalization method. The goal of this study is to ensure that the NSL data collected is on a similar scale to the other inputs. It will help develop deep learning and machine learning models that can efficiently process the collected information.

The data preprocessing process involves the preparation of the KDD NSL data for use in deep learning and machine learning algorithms. The objective of this study is to analyze the performance of the various algorithms used in deep learning and machine learning, such as the SVM, ANN, and Random Forest.

B. Feature selection

The selection of features is a crucial step in deep learning and machine learning algorithms as it directly affects their performance. It involves choosing the most relevant ones from the data and discarding irrelevant ones. Doing so helps reduce the dataset's complexity and improve its accuracy. The goal of this study is to analyze the impact of two feature selection methods on the performance of deep learning systems and machine learning techniques.

- **Principal Component Analysis (PCA)**: One of the most popular features selection techniques is the transformation of the original dataset into a set of uncorrelated variables, which are referred to as principal components. This process can help improve the performance of deep learning systems and machine learning algorithms. The method utilized in this study is PCA, which will be used to identify the most crucial features from the KDD-NSL dataset. It will then be evaluated on the performance of the ANN, SVM, and Random Forest algorithms.

C. Machine learning and deep learning algorithms used

Anomalies detection using deep learning and machine learning techniques are commonly used in network traffic data to identify anomalies. In this paper, we will introduce three different algorithms that will be used to detect anomalous patterns in the KDD network traffic data.

- Support Vector Machine (SVM): SVM is a widely used machine learning algorithm for analyzing and classifying network traffic data. It can be used to classify the data into different groups. In network anomaly detection, it is known to be effective at identifying anomalous patterns.
- Random Forest: Random Forest combines several decision trees to improve its accuracy and reduce overfitting. It can be used to detect network anomalies. In terms of its performance, Random Forest is relatively insensitive to the data's dimensionality.
- Artificial Neural Network (ANN): An artificial neural network is a type of deep learning system that is modeled after the brain's structure and function. It can be used to identify network anomalies. In this paper we will introduce three different ANN algorithms that will be used in the classification of the KDD network traffic.

The three algorithms will be evaluated using various metrics, such as accuracy, recall, and F1-point score. The results will be shown in graphs-1 and table-1,2, and we will identify the best algorithm for detecting KDD network anomalies.

DOI: https://doi.org/10.17762/turcomat.v10i2.13626

Results and Outputs

i. Without feature selection

Table 1 Evaluation without feature selection

Algorithm	Accuracy	Precision	Recall	F1-Score
SVM	92	93	89	91
Random Forest	94	96	91	93
ANN	95	94	93	93

ii. With feature selection

Table 2 Evaluation with feature selection

Algorithm	Accuracy	Precision	Recall	F1-Score
SVM	96	97	95	96
Random Forest	97	98	96	97
ANN	98	97	97	97

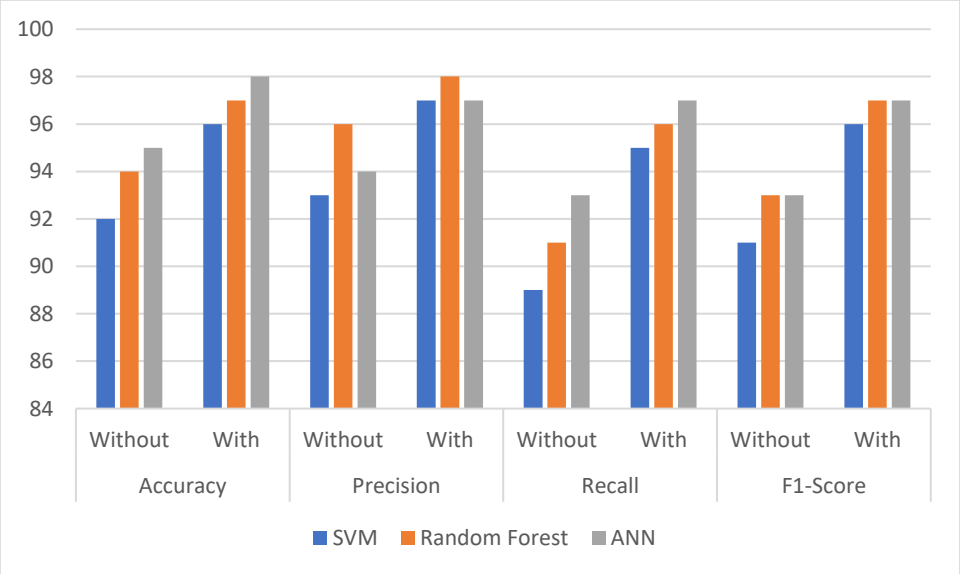

Figure 1 Comparative graph

Although the three algorithms did well when it came to detecting network anomalies without feature selection, they performed significantly better when it came to capturing the most relevant data with the selected features. The improvements in the accuracy, recall, F1-score, and precision of the three algorithms were significant. The results indicate that selecting the right features can significantly improve deep learning and machine learning algorithms' performance when detecting anomalies in networks.

Conclusion and future scope

The paper presents a study on the use of deep learning and machine learning techniques to detect anomalous events in network traffic. We utilized the KDD-NSL dataset and three popular algorithms namely, the SVM, ANN, and Random Forest. We also utilized feature selection methods to improve the performance. The results of our study revealed that the three algorithms that were used to detect network anomalies were able to perform well in terms of their accuracy, recall, F1-score, and precision. Furthermore, the selection of features led to a significant increase in the performance of the algorithms. The findings of this study show that deep learning and machine learning methods can effectively identify network anomalies. They also suggest that feature selection can help

improve the performance of these techniques. Due to the increasing volume of network data and the complexity of the situation, cyber-attacks are becoming more prevalent.

The findings of this study suggest that future research should focus on developing better algorithms that can perform better than their current counterparts. In addition, developing models that can adapt to cyber threats should be pursued. Researchers can explore the applications of deep learning and machine learning methods in detecting network anomalies. These include the use of neural networks such as CNNs and RNNs, which can perform better than traditional methods in capturing complex correlations and patterns. The study's findings provide valuable insight into the applications of machine learning and deep learning in detecting network anomalies and securing networks.

References

[1] S. H. A. H. Baddar, A. Merlo, and M. Migliardi, "Anomaly detection in computer networks: A state-of-the-art review," *J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl.*, vol. 5, no. 4, pp. 29–64, 2014.

[2] Y. Xin *et al.*, "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018, doi: 10.1109/ACCESS.2018.2836950.

[3] S. Y. Huang and Y. N. Huang, "Network traffic anomaly detection based on growing hierarchical SOM," *Proc. Int. Conf. Dependable Syst. Networks*, pp. 10–11, 2013, doi: 10.1109/DSN.2013.6575338.

[4] Z. Du, L. Ma, H. Li, Q. Li, G. Sun, and Z. Liu, "Network Traffic Anomaly Detection Based on Wavelet Analysis," *Proc. - 2018 IEEE/ACIS 16th Int. Conf. Softw. Eng. Res. Manag. Appl. SERA 2018*, pp. 94–101, 2018, doi: 10.1109/SERA.2018.8477230.

[5] O. I. Sheluhin and I. Y. Lukin, "Network Traffic Anomalies Detection Using a Fixing Method of Multifractal Dimension Jumps in a Real-Time Mode," *Autom. Control Comput. Sci.*, vol. 52, no. 5, pp. 421–430, 2018, doi: 10.3103/S0146411618050115.

[6] B. J. Radford, L. M. Apolonio, A. J. Trias, and J. A. Simpson, "Network Traffic Anomaly Detection Using Recurrent Neural Networks," pp. 1–7, 2018, [Online]. Available: http://arxiv.org/abs/1803.10769.

[7] M. Mantere, M. Sailio, and S. Noponen, "Network traffic features for anomaly detection in specific industrial control system network," *Futur. Internet*, vol. 5, no. 4, pp. 460–473, 2013, doi: 10.3390/fi5040460.

[8] M. A. A. Naser, "Network Traffic Analysis based on collective anomaly detection," pp. 1141–1146, 2014.

[9] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Mach. Learn.*, vol. 101, no. 1–3, pp. 59–84, 2015, doi: 10.1007/s10994-014-5473-9.

[10] T. Andrysiak, Ł. Saganowski, and W. Mazurczyk, "Network anomaly detection for railway critical infrastructure based on autoregressive fractional integrated moving average," *Eurasip J. Wirel. Commun. Netw.*, vol. 2016, no. 1, 2016, doi: 10.1186/s13638-016-0744-8.

[11] M. Ding and H. Tian, "PCA-based network Traffic anomaly detection," *Tsinghua Sci. Technol.*, vol. 21, no. 5, pp. 500–509, 2016, doi: 10.1109/TST.2016.7590319.

[12] L. Nie, D. Jiang, and Z. Lv, "Modeling network traffic for traffic matrix estimation and anomaly detection based on Bayesian network in cloud computing networks," *Ann. des Telecommun. Telecommun.*, vol. 72, no. 5–6, pp. 297–305, 2017, doi: 10.1007/s12243-016-0546-3.

[13] S. Naseer *et al.*, "Enhanced network anomaly detection based on deep neural networks," *IEEE Access*, vol. 6, pp. 48231–48246, 2018, doi: 10.1109/ACCESS.2018.2863036.

[14] D. H. Hoang and H. D. Nguyen, "A PCA-based method for IoT network traffic anomaly detection," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2018-February, pp. 381–386, 2018, doi: 10.23919/ICACT.2018.8323766.