Turkish Journal of Computer and Mathematics Education *Vol.10 No.02 (2019), 1075-1080*
DOI: https://doi.org/10.17762/turcomat.v10i2.13628

Research Article

Machine Learning for Network Security: An Analysis of Intrusion Detection Systems

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

Abstract- Due to the increasing importance of data and communication over computer networks, securing the network has become a critical issue. One of the most common factors that attackers use to bypass security measures is the availability of Intrusion Detection System (IDS) techniques. This paper aims to introduce machine learning to help improve the capabilities of IDS. The paper explores the potential of machine learning to enhance the security of networks. It starts by introducing network security and IDS, and then it reviews the current limitations of IDS techniques. It then delves into the world of machine learning and its applications in IDS. The paper presents an overview of the various aspects of machine learning-based IDS, including the datasets used for analysis, the models used, and the evaluation metrics that were utilized to measure their effectiveness. It then compares the results with traditional IDS techniques, and it explores the potential of this technology for future research. According to the findings of the study, machine learning-based intrusion detection systems can provide an efficient and accurate method of detecting unauthorized access to a computer network. The paper concludes by discussing the technology's potential for securing networks.
Keywords: Network security, IDS, Machine Learning, Cyber-attack

Introduction

Due to the rise of the internet and the increasing number of people using it, cybersecurity has become more important than ever. Attacks that can cause significant losses or compromise the national security of an organization or individual are now more likely to occur. To prevent these types of threats, network security experts use various measures such as intrusion detection systems. In addition to being able to monitor the traffic in the network, an intrusion detection system (IDS) also helps security teams identify potential threats. This technology is very important for network security as it allows them to respond to these threats before they become a major issue. Due to the emergence of machine learning, IDS can now detect complex attacks that would typically be missed by traditional methods.[1], [2]

Due to the increasing number of people and organizations using computer networks, cybersecurity has become more important. Cybercriminals are constantly looking for ways to attack the networks, which can be carried out through various techniques such as phishing attacks and malware. These attacks can cause a lot of damage, including financial losses and reputational damage. Several organizations, such as banks and hospitals, are vulnerable to cyberattacks. A successful attack on a financial institution could result in the loss of customer trust and damage the reputation of the bank. On the other hand, a cyberattack on a government agency could affect the national security of the country.

An intrusion detection system is a type of security device that monitors the traffic in an organization's network and informs security teams if something is wrong. It can be classified into one of two types: anomaly-based or signature-based.

- An IDS that uses a signature-based approach can analyze known threats and provide a list of possible attacks. However, it is not able to detect new and sophisticated attacks. This type of security system is mainly beneficial for organizations that have predefined rules.
- An anomaly-based IDS uses a combination of machine learning techniques to analyze the network's normal behavior and identify new threats. This type of system can also detect previously unrecognized attacks. One disadvantage of this type of security system is that it can generate false positives.

The field of machine learning is a sub-field of AI that involves learning through data. It can be used to create models that can analyze and detect anomalous and complex network traffic. Machine learning can help improve the efficiency and accuracy of an intrusion detection system by allowing it to analyze and detect previously unrecognized threats. Traditional IDS can be limited by how security experts can develop rules that can identify known threats. With machine learning-based IDS, they can learn from network traffic and identify previously unknown attacks.[3], [4]

An IDS using machine learning algorithms can be created by analyzing various types of data, such as decision trees and neural networks. These models can then be used to identify and prevent various types of attacks, such as phishing attacks and distributed denial of service (DDoS). In addition, machine learning-based systems can also be used to identify insider threats, which are usually carried out by individuals using their privileges.[5]

DOI: https://doi.org/10.17762/turcomat.v10i2.13628

Due to the increasing number of studies being conducted on the use of machine learning in an IDS, the research community has been continuously reviewing the various aspects of this technology. This paper aims to provide a comprehensive overview of the current state of the research on this technology and its potential to improve the security system's efficiency. In this paper, we present an overview of the various aspects of machine learning-based intrusion detection systems. We show that it can effectively identify security threats and perform better than signature-based IDS when it comes to detecting unknown ones. Despite the advantages of machine learning in IDS, there are still some issues that need to be resolved in order to improve its performance. One of these is the need for more robust algorithms.

Literature review

R. Kumari et al.[6] presents an anomaly detection technique for network traffic using the K-means clustering algorithm. The authors propose a method that utilizes traffic statistics and traffic data features to identify anomalous traffic. They evaluate the performance of their technique on the NSL-KDD dataset and achieve an accuracy of 99.03%.

Y. Xin et al. [7] provides an overview of machine learning and deep learning methods for cybersecurity, including intrusion detection systems (IDS). The authors discuss the application of various machine learning algorithms in IDS and highlight their advantages and limitations. They also discuss the use of deep learning models in IDS and provide recommendations for future research.

M. Ahmed et al.[8] propose a novel approach for network traffic pattern analysis using clustering-based collective anomaly detection. They use a clustering algorithm to group similar network connections together and detect anomalies based on deviations from the typical behavior of these clusters. The authors evaluate their approach on the KDD Cup 99 dataset and achieve an accuracy of 99.99%.

F. Iglesias et al.[9] analyzes network traffic features for anomaly detection and proposes a method that combines multiple classifiers to improve detection accuracy. The authors evaluate their approach on the KDD Cup 99 dataset and achieve an accuracy of 99.94%. They also compare their approach with other anomaly detection methods and discuss the advantages and limitations of each approach.

Huang et al.[10] proposed an anomaly detection technique based on the growing hierarchical self-organizing map (GHSOM) for network traffic. The GHSOM algorithm is utilized to analyze and identify the different traffic patterns in the network traffic data, and the differences between normal and anomalous traffic patterns are measured using the Euclidean distance metric. The proposed approach was tested on the NSL-KDD dataset, and the results show that it achieved high accuracy in detecting network traffic anomalies.

Sheluhin et al.[11] proposed a novel method for detecting network traffic anomalies in real-time using multifractal dimension jumps. The method involves detecting changes in the multifractal dimensions of network traffic data streams and then identifying these changes as network anomalies. The proposed method was tested on a real network traffic dataset and achieved high accuracy in detecting network anomalies.

Atli et al.[12] proposed an anomaly-based intrusion detection approach that combines extreme learning machine (ELM) and network traffic statistics aggregation in probability space. The ELM algorithm is used to classify normal and anomalous traffic, and the aggregation of network traffic statistics in probability space is used to model the statistical behavior of network traffic. The proposed approach was tested on the KDDCup'99 dataset and achieved high accuracy in detecting network intrusions.

Ahmed et al.[13] presents a survey of collective anomaly detection techniques for network traffic analysis. The paper provides an overview of various techniques for detecting network anomalies, including statistical methods, clustering-based methods, and machine learning-based methods. The paper also presents a comparison of these techniques based on their performance in detecting network anomalies.

Radford et al.[14] proposed an anomaly detection approach based on recurrent neural networks (RNNs) for network traffic analysis. The proposed approach utilizes the long short-term memory (LSTM) architecture of RNNs to capture the temporal dependencies of network traffic data. The approach was tested on the NSL-KDD dataset and achieved high accuracy in detecting network traffic anomalies.

Mao et al.[3] provide a comprehensive survey of deep learning techniques for intelligent wireless networks. While the paper covers a wide range of topics, it includes a section on intrusion detection systems and discusses the potential of deep learning methods for improving the accuracy of such systems.

Network security is of paramount importance in today's digital landscape, and intrusion detection systems play a crucial role in ensuring the integrity and availability of network resources. While traditional intrusion detection systems have limitations in detecting novel and sophisticated attacks, machine learning-based approaches show promise in improving the accuracy and effectiveness of intrusion detection systems. Several studies have proposed

DOI: https://doi.org/10.17762/turcomat.v10i2.13628

and evaluated different machine learning techniques for intrusion detection on the NSL-KDD dataset, with SVM, Naive Bayes, Decision Tree, and KNN being the most commonly used algorithms. However, there is still a need for further research and development to address the challenges and limitations of machine learning-based intrusion detection systems.

Difference between Traditional approach vs Machine Learning based approach for IDS

An IDS is a vital component of network security, as it can identify potential threats. Traditional IDS techniques utilize a set of rules and signatures to define attack patterns, which can then be used to trigger alarms if they match. However, this approach can be very inefficient when it comes to detecting attacks that have not been identified.

In contrast, an IDS that uses machine learning techniques can learn and adapt to different attack patterns. It can then analyze network traffic and determine which patterns are malicious or normal. This method can also detect complex attack techniques that the traditional IDS cannot identify.

There are significant differences between an IDS using machine learning techniques and a traditional one.

- Approach: The traditional approach of an IDS is to use a set of predefined rules or signatures to identify potential threats. In contrast, a machine learning-based one uses data-driven techniques to analyze network traffic and identify patterns that indicate malicious or normal behavior.
- Performance: The main advantage of a traditional IDS is its ability to detect known threats, but it can also be very inefficient when it comes to identifying attacks that are not related to the signature or rules. With the use of machine learning techniques, an IDS can now identify complex attack methods that it may not have been able to detect before.
- Scalability: The scalability of a traditional IDS is limited by the need for updates to the signatures or rules to stay effective. With the use of machine-learning techniques, an IDS can learn to identify new attack patterns and threats.
- False Positives: If the IDS's rules or signatures are not well-defined, it can generate false positives. With the help of machine learning, an IDS can learn about the network's normal behavior and identify anomalies that would not have been detected with traditional techniques.
- Training Data: Traditional IDS require relatively little training data to operate properly, as they only rely on predefined rules and signatures. On the other hand, machine learning-based systems require large amounts of training data to identify patterns related to malicious or normal activities.

Although a traditional IDS can effectively identify known threats, it can also be inefficient when it comes to finding attacks that do not match their signatures or rules. With the help of machine learning, an IDS can learn to identify patterns related to malicious or normal behavior. Unfortunately, this method requires a large amount of data to train effectively, and it can be vulnerable to attacks.

Methodology

i. Dataset

This paper aims to analyze the performance of different machine learning algorithms when it comes to detecting intrusions. We utilize the NSL-KPDD dataset, which is used in the field of intrusion detection. The dataset is a modified variation of the KDD Cup. The NSL-KDD network traffic data set contains a set of network traffic information that has either been labeled as malicious or normal traffic. It has a total of 41 features.

ii. Pre-processing

Many organizations use the NSL–KDD dataset for their intrusion detection needs. Preprocessing the data is an essential step prior to training a machine learning model, and we will cover the three steps we utilized for this project.

a. Data Cleaning: The data cleaning process is carried out to remove or correct data that is inaccurate, incomplete, and irrelevant. In the KDD-NSL dataset, there are numerous redundant features and duplicates. We also removed some features that have no significant impact on the classification process.
b. Feature Scaling: The process of scaling a feature to a similar range is known as feature scaling. It is important for machine learning algorithms to be sensitive to the varying ranges of inputs in the NSL-KDD dataset. To minimize variance and mean, we used the Standard Scaler library.
c. Feature Selection: The feature selection process is carried out to identify the subsets of the data that are relevant to the classification task. It can help improve the training process and reduce the overall dimensionality of the dataset. In the NSL KDD dataset, there are over 40 features, but only a few of them are relevant to the classification task.

DOI: https://doi.org/10.17762/turcomat.v10i2.13628

We performed three steps in the preparation of the NSL-KPDD dataset. These included data cleaning, feature selection, and scaling. These steps helped us improve the performance of the training models and prepare the dataset for further analysis.

iii. ML Algorithm used

In this section, we will talk about the various algorithms that are used to detect intrusions in the NSL-KDD database. These include the Naive Bayes, SVM, Decision Tree, and KNN.

a. SVM - SVM is a widely used machine learning algorithm that is used for classification tasks. It seeks to find the best hyperplane for separating the various classes of data into their respective categories. It can perform well with both non-linear and linearly skewed datasets.

b. Naïve Bayes - The Naive Bayes algorithm is a simple and effective method for classification. It assumes that the various features of a given set are independent of one another and then produces a conditional probability for each class. It is suitable for large datasets due to its fast performance and small amount of memory. However, its accuracy can be affected by the assumption of independence.

c. Decision Tree - A decision tree is a representation of various decisions that are related to a given set of data. It can be created by an algorithm that recursively splits the data into various classes. Although it is easy to interpret, decision trees can get overwhelmed by the data and may not perform well with new information.

d. KNN - Classification is done by implementing the KNN algorithm, which takes into account the data points that are most commonly found in the training set and classifies them into a class that is most likely to perform well. It is simple to implement and can work well with large datasets, but it can be very expensive due to its sensitivity to k selection.

The SVM algorithm is powerful enough to handle high-dimensional data. Naive Bayes requires little memory and is fast, while KNN is inexpensive and can be used for small datasets. Each of these algorithms has its weaknesses and strengths, and the decision on which to implement depends on the security system's goals and the characteristics of the data.

Results and Output

Table 1 Evaluation of NSL-KDD with various ML algo

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
SVM	98.91	98.89	98.83	98.86	99.61
Naive Bayes	78.54	81.05	74.26	77.52	86.44
Decision Tree	99.57	99.62	99.56	99.59	99.61
KNN	97.87	97.87	97.86	97.87	99.18

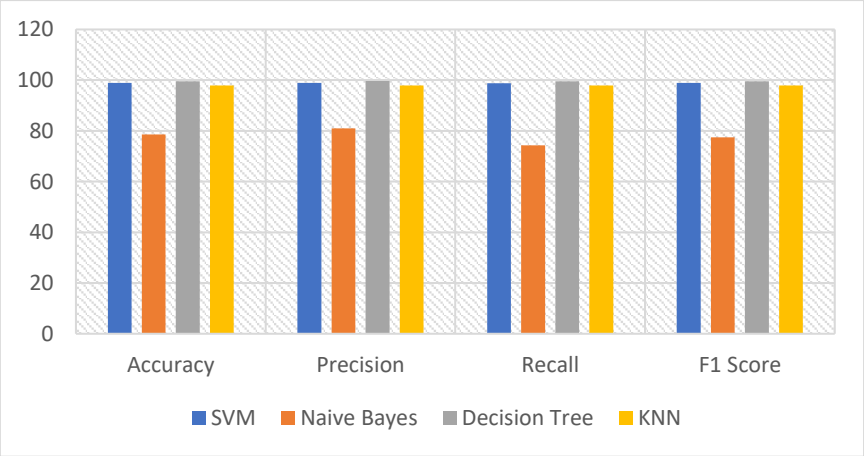

Figure 1 Graphical representation of evaluation parameters

DOI: https://doi.org/10.17762/turcomat.v10i2.13628

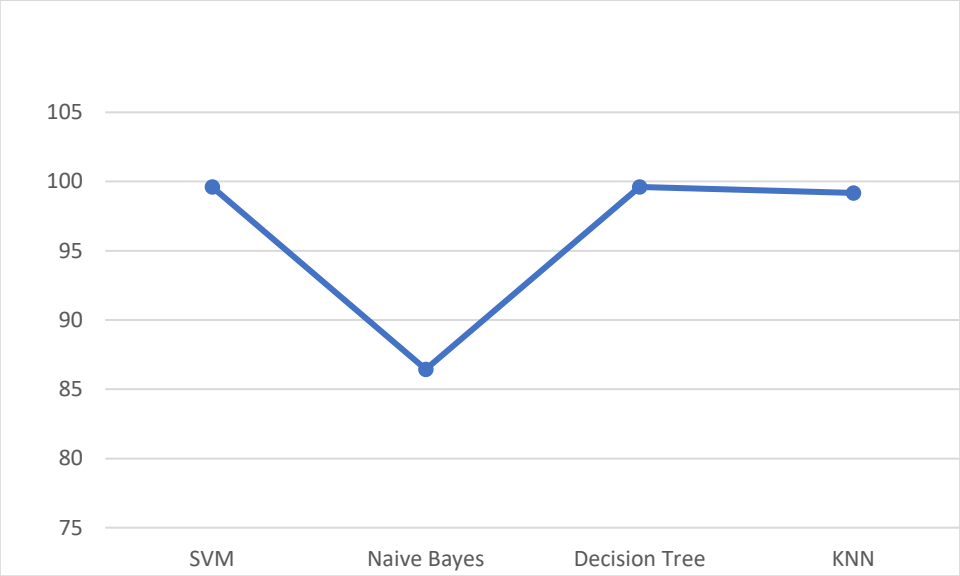

Figure 2 Graphical representation of AUC

The table-1 and figure 1,2 summarizes the various evaluation metrics used includes accuracy, recall, precision, AUC, and F1 score. With regard to accuracy, it measures how accurately the system can classify instances. On the other hand, with regard to precision, it measures how accurately the system can predict true positives. The F1 score is a harmonic representation of recall and precision. The AUC is a statistical measure of the difference between false positive and true positive rates. The performance of various machine learning algorithms when it comes to detecting intrusions depends on the type of dataset it is dealing with and its goals. For instance, the SVM and Decision Tree are powerful models that perform well on the NSL KDD dataset. On the other hand, the Naive Bayes algorithm has the lowest performance.

Conclusion and Future scope

The paper presents an analysis of the performance of various machine learning algorithms in detecting intrusions using a dataset known as the NSL–KDD. The findings show that the Decision Tree and SVM algorithms perform well while the Naïve Bayes algorithm is the worst performer. These findings support the idea that machine learning could be a valuable tool in improving network security. Due to the increasing sophistication and frequency of cyber-attacks, network security has become a critical component of today's digital landscape. An effective intrusion detection system is a must-have for protecting networks. Traditional systems can't provide effective and accurate intrusion detection. With the help of machine learning algorithms, they can be improved by detecting anomalous network traffic and learning from data. The paper shows that machine learning can be used in the detection of intrusions. The researchers found that the Decision Trees and SVM algorithms performed well in terms of their accuracy, recall, AUC-ROC, and F1 score. They can be utilized for various scenarios and datasets, which shows that they can be a valuable tool for network security. The paper provides an overview of the various applications of machine learning in detecting intrusions. It also explores the possible applications of this technology in improving network security.

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

[3] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 2595–2621, 2018, doi: 10.1109/COMST.2018.2846401.

[4] 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.

[5] 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-Febru, pp. 381–386, 2018, doi:

10.23919/ICACT.2018.8323766.

[6] R. Kumari, Sheetanshu, M. K. Singh, R. Jha, and N. K. Singh, "Anomaly detection in network traffic using K-mean clustering," *2016 3rd Int. Conf. Recent Adv. Inf. Technol. RAIT 2016*, pp. 387–393, 2016, doi: 10.1109/RAIT.2016.7507933.

[7] 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.

[8] M. Ahmed and A. N. Mahmood, "Novel Approach for Network Traffic Pattern Analysis using Clustering-based Collective Anomaly Detection," *Ann. Data Sci.*, vol. 2, no. 1, pp. 111–130, 2015, doi: 10.1007/s40745-015-0035-y.

[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] 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.

[11] 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.

[12] B. G. Atli, Y. Miche, A. Kalliola, I. Oliver, S. Holtmanns, and A. Lendasse, "Anomaly-Based Intrusion Detection Using Extreme Learning Machine and Aggregation of Network Traffic Statistics in Probability Space," *Cognit. Comput.*, vol. 10, no. 5, pp. 848–863, 2018, doi: 10.1007/s12559-018-9564-y.

[13] M. Ahmed, "Collective Anomaly Detection Techniques for Network Traffic Analysis," *Ann. Data Sci.*, vol. 5, no. 4, pp. 497–512, 2018, doi: 10.1007/s40745-018-0149-0.

[14] 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.