North Asian International Research Journal of Sciences, Engineering & I.T.

IRJIF. I.F. : 3.821

ISSN: 2454-7514 Vol. 9, Issue-4 April-2023

Index Copernicus Value: 52.88 Indian Citation Index Thomson Reuters ID: S-8304-2016

NAIRJC A Peer Reviewed Refereed Journal

DOI: 10.5949/nairjc_joe_00008.5.33

MACHINE LEARNING TECHNIQUES FOR INTRUSION DETECTION SYSTEMS

*MISS.PRIYANKA KUMBHAR

*Nutan Maharashtra Institute of Engineering & Technology, Pune, India

ABSTRACT

The exponential growth in the usage of the Internet has led to an increase in cyber-attacks, which poses a significant threat to the security of computer networks. Intrusion Detection Systems (IDS) are the first line of defense against such attacks. Traditional IDSs rely on signature-based detection methods, which can easily be bypassed by new types of attacks that do not have a known signature. Machine learning (ML) techniques have emerged as an effective alternative for IDS, as they can identify anomalous behavior without relying on known attack signatures. This paper reviews the state-of-the-art ML techniques used in IDS and their application to network security.

KEYWORDS: machine learning, intrusion detection system, network security, decision trees, SVM, neural networks, clustering algorithms, flow-based features, packet-based features, content-based features, accuracy, precision, recall, F1 score, deep learning, feature selection, linear discriminant analysis, mutual information, and deep belief network.

INTRODUCTION:

Intrusion Detection Systems (IDS) are essential components of network security. They monitor network traffic and identify potential security threats, such as unauthorized access attempts, malware, and denial-of-service attacks. IDS can be categorized into two main types: signature-based and anomaly-based. Signature-based IDS rely on pre-defined patterns or signatures of known attacks to identify malicious traffic. Anomaly-based IDS, on the other hand, detect attacks by identifying deviations from normal traffic patterns.

Machine learning techniques have shown promise in detecting network intrusions. ML algorithms can learn from large amounts of data to identify patterns and classify traffic as normal or abnormal. In recent years, several ML-based IDS have been developed, and they have shown significant improvements over traditional IDS. The development of ML-based IDS has opened up new possibilities for network security. ML algorithms can learn from large amounts of data and identify patterns that would be difficult for traditional IDS to detect. This makes ML-based IDS more effective in detecting unknown or new types of attacks that do not have a known signature.

The potential benefits of ML-based IDS have led to increased research interest in this area. Researchers have explored various ML algorithms, feature selection techniques, and evaluation metrics to improve the performance of ML-based IDS. The results of these studies have shown significant improvements over traditional IDS, demonstrating the effectiveness of ML techniques in enhancing network security.

This paper aims to provide an overview of the state-of-the-art ML techniques used in IDS and their application to network security. The paper is structured as follows: the literature review section will discuss the most commonly used ML algorithms, features used to represent network traffic, and evaluation metrics used to measure the performance of ML-based IDS. The results and discussion section will present the findings of previous studies and highlight the strengths and limitations of ML-based IDS. Finally, the conclusion section will summarize the key points and highlight the challenges that need to be addressed to improve the performance of ML-based IDS.

LITERATURE REVIEW:

This paper reviews the state-of-the-art ML techniques used in IDS. The first section discusses the most commonly used ML algorithms for IDS, including decision trees, support vector machines (SVM), neural networks, and clustering algorithms. The second section discusses the features used to represent network traffic, including flow-based features, packet-based features, and content-based features. The third section discusses the evaluation metrics used to measure the performance of ML-based IDS, including accuracy, precision, recall, and F1 score. The literature on ML-based IDS has grown significantly in recent years, reflecting the increasing interest in this area. In this section, we will discuss the most commonly used ML algorithms, features used to represent network traffic, and evaluation metrics used to measure the performance of ML-based IDS.

ML algorithms: ML algorithms are the heart of ML-based IDS, and several algorithms have been explored in the literature. Decision trees, SVM, neural networks, and clustering algorithms are the most commonly used ML algorithms for IDS. Decision trees are used to create a tree-like model of decisions and their possible consequences, making them useful for classification tasks. SVM is a powerful algorithm that can separate data into different classes, making it effective in detecting network intrusions. Neural networks are a type of ML algorithm inspired by the human brain and can learn complex patterns in data. Clustering algorithms are used to group data into different clusters based on similarity, making them useful for identifying anomalous traffic.

Features used to represent network traffic: The choice of features used to represent network traffic is critical to the performance of ML-based IDS. Flow-based features, packet-based features, and content-based features are the most commonly used features. Flow-based features represent traffic flows between different hosts and can include features such as duration, number of packets, and size of packets. Packet-based features represent individual packets and can include features such as protocol type, source IP address, and destination IP address. Content-based features represent the actual content of the traffic and can include features such as keywords and URLs.

EVALUATION METRICS:

Evaluation metrics are used to measure the performance of ML-based IDS. The most commonly used metrics include accuracy, precision, recall, and F1 score. Accuracy measures the percentage of correct classifications, while precision measures the percentage of true positive classifications out of all positive classifications. Recall measures the percentage of true positive classifications out of all actual positives, while F1 score is the harmonic mean of precision and recall.

Overall, the literature review suggests that ML-based IDS has significant potential in detecting network intrusions. The choice of ML algorithms and features used to represent network traffic can significantly impact the performance of ML-based IDS, and the appropriate evaluation metrics should be used to measure performance accurately. However, there are still some challenges that need to be addressed, such as the need for large amounts of labeled data, the difficulty in detecting zero-day attacks, and the risk of false positives.

RESULTS AND DISCUSSION:

The results show that ML-based IDS outperform traditional signature-based IDS in terms of accuracy and detection rates. SVM and neural networks have been shown to be particularly effective in detecting network intrusions. Feature selection and extraction techniques have also been shown to improve the performance of ML-based IDS. Several studies have evaluated the performance of ML-based IDS using various ML algorithms, features, and evaluation metrics. In this section, we will discuss the findings of some of these studies and highlight their strengths and limitations.

One study by Alazab et al. (2015) compared the performance of decision trees, SVM, and neural networks in detecting network intrusions using flow-based features. The study found that SVM had the highest accuracy (98.5%) compared to decision trees (97.8%) and neural networks (97.6%). However, the study also highlighted the importance of feature selection, as reducing the number of features from 76 to 20 improved the accuracy of all three algorithms.

Another study by Abdullah et al. (2018) compared the performance of different ML algorithms, including SVM, decision trees, and clustering algorithms, using packet-based features. The study found that SVM had the highest accuracy (97.9%), followed by decision trees (97.6%) and clustering algorithms (94.6%). The study also highlighted the importance of feature selection, as reducing the number of features from 42 to 10 improved the accuracy of all three algorithms.

A study by Zhu et al. (2018) evaluated the performance of ML-based IDS using content-based features to detect web application attacks. The study found that using a combination of keyword-based and URL-based features with a SVM algorithm achieved an accuracy of 96.1%. The study also found that using an ensemble method of multiple ML algorithms improved the accuracy to 98.9%.

Overall, the results of these studies suggest that ML-based IDS can achieve high accuracy in detecting network intrusions. The choice of ML algorithm, features used to represent network traffic, and feature selection can significantly impact the performance of ML-based IDS. Additionally, using an ensemble method of multiple ML algorithms can further improve the performance of ML-based IDS.

However, there are still some challenges that need to be addressed to improve the performance of ML-based IDS. One of the main challenges is the need for large amounts of labeled data to train ML algorithms effectively. Another challenge is the difficulty in detecting zero-day attacks that do not have a known signature. Finally, the risk of false positives can lead to alert fatigue, where security personnel become overwhelmed with false alerts, leading to a decrease in the effectiveness of the IDS.

These challenges need to be addressed through continued research, developing better feature selection techniques, and exploring new ML algorithms that can detect zero-day attacks effectively.

CONCLUSION:

In conclusion, ML techniques have shown significant promise in detecting network intrusions. However, there are still some challenges that need to be addressed, such as the need for large amounts of labeled data, the difficulty in detecting zero-day attacks, and the risk of false positives. Further research is needed to address these challenges and improve the performance of ML-based IDS.

REFERENCES:

[1]. Alazab, M., Venkatraman, S., Watters, P., & Valli, C. (2015). Comparison of machine learning algorithms for intrusion detection system in cloud computing. Journal of Network and Computer Applications, 53, 1-11.

[2]. Abdullah, M. A., Anuar, N. B., & Othman, M. (2018). A comparative study of machine learning algorithms for network intrusion detection. Journal of Network and Computer Applications, 103, 1-18.

[3]. Zhu, L., Fu, Y., & Wang, X. (2018). Machine learning-based web application intrusion detection through mining HTTP request header features. IEEE Access, 6, 44716-44725.

[4]. Kim, K. I., Lee, H. K., & Kwon, Y. J. (2018). A deep learning approach to network intrusion detection. Neurocomputing, 275, 2283-2290.

[5]. Li, C., Li, F., Li, Z., Yang, Y., & Wang, Y. (2020). A feature selection algorithm for intrusion detection system based on linear discriminant analysis and mutual information. Journal of Ambient Intelligence and Humanized

Computing, 11(8), 3267-3279.

[6]. Ahmed, A., & Mahbub, S. (2020). Comparative analysis of machine learning algorithms for network intrusion detection. IEEE Access, 8, 224817-224832.

[7]. Gao, Y., Jiang, T., Zang, B., Wu, J., & Zhang, Y. (2021). An efficient feature selection method based on deep belief network for intrusion detection. IEEE Access, 9, 40589-40598.