International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 9, ISSN No. 2455-2143, Pages 225-230
Published Online January 2022 in IJEAST (http://www.ijeast.com)

A SURVEY ON NETWORK-BASED INTRUSION DETECTION SYSTEMS USING MACHINE LEARNING ALGORITHMS

Amin Lama
Student
Department of MCA
JAIN (Deemed-to-be University)
Bangalore, India

Dr. Preeti Savant (Project Guide)
Professor
Department of MCA
JAIN (Deemed-to-be University)
Bangalore, India

***Abstract*—Network security is of central significance in the current information world. Due to the rapid increase of network-enabled devices, there is a significant risk of network intrusion more than ever. Hackers and intruders can successfully attack to cause the crash of the networks and web services by the unauthorized intrusion, which may cause a significant loss to an organization in terms of data and money. So, it is high time to create an intrusion detection system that can detect all types of intrusion. Due to the rapid growth and significant results of machine learning (ML) algorithms in several areas, there has recently been much interest in applying them to network security. The network-based intrusion detection system (NIDS) has much promise to be the borderline of defense against intrusions in the current information communication technology (ICT) era, and it's a critical aspect of network security. Due to the dynamic nature of attacks, intrusion detection datasets are available publicly. Intrusion detection systems are the backbone of the networks and data protection. Various IDS approaches have been used over time to achieve maximum detection accuracy. This paper investigates the different machine learning methods used to deploy network-based intrusion detection systems. This survey could give scholars a better grasp of present methodologies and help them find research possibilities and do further research in this area.**

***Keywords*—intrusion detection system, machine learning, network-based intrusion detection system, cyberattacks, network security.**

I. INTRODUCTION

An Intrusion Detection System (IDS) is used to detect network intrusion. It monitors network traffic for malicious activity or policy violations and alerts when such action is discovered. The intrusion detection system can be classified into two categories:

- Network-Based IDS: It identifies intrusions by analyzing incoming network traffic.
- Host-Based IDS: It identifies intrusions by analyzing system logs, file-system modifications, and system calls.

Intrusion Detection techniques can also be classified into two categories:

- Anomaly Detection: It detects malicious traffic by looking for variations from conventional network traffic patterns. It identifies anomalous system activity.
- Signature-Based Detection: It identifies intrusions based on previously identified patterns for malicious activity. These known patterns are called signatures.

The number of network-enabled devices connected to the internet is growing in scale and accessibility due to significant advancements in information technology. Hence, Network Security takes a pivotal role in society. However, as the number of networked devices grows, so does the potential for a more significant surface assault. As a result, global cybercrime expenses are expected to increase by 15% each year over the next five years, reaching $10.5 trillion annually by 2025, up from $3 trillion in 2015 [1]. So, it is crucial to construct a robust intrusion detection system (IDS) that combats unauthorized access to network resources to secure information [2]. Therefore, building dependable networks has become a fundamental assignment for IT administrators. However, numerous intrusion detection systems deployed today have critical flaws. They can identify the most widely recognized attack pattern using signature-based detection techniques but have the disadvantage of detecting novel attack types.To overcome this limit, many research is happening on intrusion detection system using machine learning for more dynamic approach to detect anomaly in the network.

II. DATASETS

The performance of intrusion detection systems depends on accuracy. As a result, intrusion detection accuracy must be improved to decrease false alarms and raise detection rates [3]. For this, we'll need a large dataset with both standard and abnormal behaviors. Unfortunately, the datasets used for network analysis in commercial products are not available

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 9, ISSN No. 2455-2143, Pages 225-230
Published Online January 2022 in IJEAST (http://www.ijeast.com)

publicly because of privacy considerations. However, a few datasets are open to the general public, such as KDD Cup 1999, NSL-KDD, DARPA, Kyoto 2006+, ISCX 2012, CIDDS-001, CIC-IDS2017, and CSE-CIC-IDS2018 on AWS.
KDD'99 is the older benchmark dataset that has been regularly used to evaluate NIDS performance. KDD'99 Dataset contains 4898431 instances with 41 attributes/features. It is the most used dataset even though it's old and still considered a benchmark dataset to aid academics in evaluating various intrusion detection algorithms. However, several studies suggest that evaluating a NIDS using these data sets does not reflect realistic output performance. The main problem in KDD'99 Dataset is the considerable amount of duplicate packets, which is solved by the NSL-KDD Dataset [4]. As a refined and cleaned-up version of the KDD'99, the NSL-KDD data set was created. This dataset, however, does not provide a complete picture of a modern low-footprint attack scenario.
CSE-CIC-IDS2018 on AWS is the most recent intrusion detection dataset, which is big data, publicly available, and covers a broad spectrum of attack types [5]. The Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC) collaborated on this dataset. It was created by capturing all network traffic during ten days of operation inside a controlled network environment on AWS where realistic background traffic and different attack scenarios were conducted [6]. As a result, the dataset contains about 16,000,000 instances collected over ten days [5], both benign network traffic and captures of the most common network attacks. Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and network infiltration from within are among the seven attack scenarios included in the dataset. It consists of raw PCAP network captures and processed CSV files made using CICFlowMeter-V3 that provide 80 statistical aspects of individual network flows together with their labels.

III. NETWORK-BASED INTRUSION DETECTION USING MACHINE LEARNING ALGORITHMS

An anomaly-based intrusion detection system (IDS) has been implemented to combat the increasing proliferation of new attacks. This type employs machine learning algorithms that learn from data to evaluate network traffic and determine whether it is benign or malicious [6]. Anomaly-based IDS can identify known and unknown attacks without the need for human participation or explicit scripting. The most common machine learning algorithms for NIDS implementation are:

1. Logistic Regression (LR)

Logistic Regression (LR) aims to find the best-fitting model to establish a relationship of dependence between the class variable and the features. For example, a test case with only two classes: 0 and 1, predicts a value between 0 and 1, representing the class's probability of 1 for a given observation. The simple LR model is only suitable for binary classification but can be extended for multi-class classification with some effort. Logistic regression is used in NIDS to classify traffic as an attack or normal. But to classifier the attacks into different types of attacks, multi-class logistic regression can be used. The formula of the sigmoid function is:

$$F(x) = \frac{1}{1 + e^{-(x)}}$$

Where F(x) is output between 0 and 1, x is an input to the function, and e is a base of the natural log.

2. Naïve Bayes

A Naive Bayes classifier is a probabilistic machine learning model [7] based on the Bayes rule, which assumes that the attributes are all independent of one another. The study in [8] evaluated the conditions under which the naive Bayes classifier performed better and discovered that errors were caused by three factors: noise in the training data, variance, and bias. The noise could be easily reduced by selecting appropriate training data, which a machine learning algorithm would then classify. Variance is the error caused by those too-small groupings, and bias is the error caused by huge groupings in the training data.

3. K-Nearest Neighbors (KNN)

KNN is a supervised classifier, which doesn't depend on a trained model. It is based on the assumption that an instance is most similar to the classification of other closet instances in the n-dimensional space. The target variable's output is predicted by calculating the Euclidean Distance and finding the k closest neighbors. The study in [9] proposed a classification model for NIDS based on KNN. The study used the ISCX dataset and got an accuracy of 99.96%.

4. Decision Tree (DT)

The decision tree uses labeled (pre-classified) datasets to build a tree-like classification model. It continuously divides the data into smaller structures based on specific attribute values. The root node, internal nodes, branches, and leaf nodes make up the constructed tree. The root node represents the decision tree's starting point, internal nodes represent dataset attributes, branches represent decision rules, and each leaf node indicates the final decision. To accurately split the data into appropriate classes, we must decide which attributes to include in the root node and other internal nodes. A decision tree is used for intrusion detection to classify network data flows as normal or malicious traffic; it can handle large amounts of data and detect intrusions in real-time.

5. Random Forest (RF)

The random forest algorithm is a complex nonlinear supervised algorithm used for classification and regression. This will build many decision trees when training the model, and the predictions from all trees are pooled to produce a

result. The RF classifiers work as follows: the more trees in the model, the higher the accuracy and the less the model over-fits. In NIDS, random forest uses each tree output to predict the data class (normal or attack) and then produces the final prediction by selecting the class with the most votes.

6. Support Vector Machine (SVM)

SVM is used for both classification and regression modeling. The algorithm can be trained using labeled data, and it can output the hyperplane separation of data into classes that maximize the margin among all attack classes. It blots the n-dimensional training dataset (n is the number of attributes in the dataset). The hyper-plane, which separates the classes with as wide a gap as possible, performs the classification. SVM is used in intrusion detection to create a trained model that can accurately predict the class label of each given data.

7. Gradient Boosting

Gradient boosting is a classification and regression machine learning technique. Extreme gradient boosting (XGBoost) is a type of gradient boosting that has been optimized to be highly effective and adaptable. It has recently been used to construct highly fast and accurate classification NIDS models. Alzahrani et al. [10]proposed an XGboost-based classification model for NIDS in software-defined networks. The study [10] showed that the proposed XGBoost model outperformed more than seven algorithms used in NIDS while using six different evaluation metrics.

IV. RELATED WORK

Anish Halimaa A. and K. Sundarakantham in [3] presented a study demonstrating various machine learning techniques and statistical methodologies to build an Intrusion Detection System. Machine learning techniques like Support Vector Machine (SVM) and Naïve Bayes were chosen. NSL-KDD knowledge discovery Dataset was used to evaluate the intrusion detection system. Non-numeric and symbolic features are eliminated or replaced, and instances were labeled under four categories: Normal, DoS, Probe, and R2L in pre-processing phase. Nineteen thousand instances were used for analyzing accuracy rate and misclassification rate. SVM attained an accuracy of 93.95%, while Naïve Bayes attained an accuracy rate of 71.001%, and Naïve Bayes had a high misclassification rate of 28.998%, while SVM had a misclassification rate of 2.705%. The study showed that SVM outperformed Naïve Bayes while dealing with 19,000 instances.

Chuanlong Yin et al. [11] looked into using deep learning to model an intrusion detection system. The Recurrent Neural Network (RNN) approach was used for intrusion detection. They compared it to J48, ANN, RF, SVM, and other machine learning approaches proposed by the previous researcher on the NSL-KDD dataset. The recurrent neural network IDS was well-suited to constructing a high-accuracy classification model, and its performance in binary and multi-class classification was superior to that of typical machine learning classification methods. The recurrent neural network IDS model enhanced the accuracy of intrusion detection and provided a novel research approach.

MehrnooshMonshizadeh et al. [12] present a Hybrid Anomaly Detection Model (HADM) that filters network traffic and identify malicious activities on the network. The platform applied data mining techniques to tackle the security issues effectively in high-load communication networks. They employed a protocol analyzer and a combination of linear and learning algorithms. The learning algorithms use these attributes and features to identify new forms of cyber-attacks, while the linear algorithms filter and extract distinguishing qualities and features of the cyber-attacks. For this, they used five datasets with varying sizes and attacks, including ISCX-2012, UNSW-NB15 Jan, UNSW-NB15 Feb, ISCX-2017, and MAWILab-2018. The feature selection approaches were chosen depending on the algorithm's computation time and detection rate. As a result, the HADM demonstrated robustness and scalability when tested against datasets of various sizes and assaults.

Iram et al. in [2] proposed empirical research on machine learning classifiers based on support vector machine, k-nearest neighbor, logistic regression, naïve bayes, multi-layer perceptron, random forest, decision tree, and extra-tree classifier for the classification of network data as abnormal and usual. The research performance was evaluated on four different subsets derived from the NSL-KDD dataset. Before training the model, the training data was preprocessed based on significant features. The results reveal that the machine learning classifiers produce better results for DoS attacks, and low results were achieved for U2R attacks in general accuracy of the model is 99% while using random forest, extra-tree, and decision tree classifiers.

Zena Khalid Ibrahim and Mohammed YounisThanon[13] experimented on three machine learning algorithms; random forest, k-nearest neighbor, and support vector machine implemented on the NSL-KDD dataset for the IDS system. They use a combination of recursive feature elimination (RFE) and analysis of variance (ANOVA) for feature selection in DoS attack detection. The system achieves higher accuracy using random forest using 13 features.

Sharaz Naseer et al. [14] developed anomaly detection models based on different deep neural network structures, including convolutional neural networks, autoencoder neural networks, and recurrent neural networks. NSL-KDD dataset was used to train the deep model. Well-known classification approaches, such as extreme learning machine, nearest neighbor, decision tree, random forest, support vector machine, naive-bays, and

quadratic discriminant analysis, were used to create traditional machine learning-based intrusion detection models. In addition, well-known categorization measures, such as receiver operating characteristics, were used to evaluate deep and traditional machine learning models, an area under a curve, precision-recall curve, mean average precision, and classification accuracy. Deep IDS models' experimental findings showed promise for real-world use in anomaly detection systems.

Ren et al. [15] proposed a hybrid data optimization system with two parts: data sampling and feature selection, for an effective intrusion detection system. The Isolation Forest (iForest) was used to remove outliers, the Genetic Algorithm (GA) was utilized to improve the sampling ratio, and the Random Forest (RF) classifier was used as the evaluation criteria to acquire the best training dataset. The Genetic Algorithm and Random Forest were employed to find the best feature subset in feature selection. Finally, an intrusion detection system based on the random forest was created to utilize the ideal training dataset obtained through data sampling and the features picked through feature selection. The UNSW-NB15 dataset was used to test the model, which revealed that it outperformed other algorithms in recognizing unusual anomaly behaviors.

AbdulsalamAlzahrani et al. [10] presented a study demonstrating the use of machine learning methods for traffic monitoring as part of a network intrusion detection system in the software-defined networks controller to detect malicious activities in the network. Three different tree-based machine learning approaches, Decision Tree, Random Forest, and XGBoost, were used to show attack detection. The proposed methods were trained and tested using the NSL-KDD dataset. Many complex preprocessing techniques were performed to obtain the optimum form of data. Only 5 of the NSL-KDD dataset's 41 features were used, and a multi-class classification task was completed by detecting whether an attack had occurred and classified the type of attack as DDoS, Probe, R2L, or U2R. The study showed an accuracy of 95.55%.

V. Kanimozhi and Prem Jacob [16] proposed a system created by applying artificial intelligence on a realistic cyber defense dataset CSE-CIC-IDS2018 on AWS. The proposed system of Artificial Neural Networks shows an outstanding Accuracy Score of 99.97% and an average area under the Receiver Operator Characteristic (ROC) curve was 0.999, and an average False Positive rate was a mere value of 0.001. The botnet attack detection was powerful, accurate, and precise on the proposed system. The proposed system can be implemented in n machines to analyze conventional network traffic, cyber-physical system traffic, and real-time network traffic.

Gisung Kim et al. [17] present a hybrid intrusion detection method that combines misuse detection and anomaly detection in a decomposed structure hierarchically. The C4.5 decision tree was used to develop the misuse detection model, which was then used to break down the normal training data into smaller subgroups. The anomaly detection model is created in each subdivided region using a one-class support vector machine (1-class SVM). The anomaly detection model might indirectly leverage known attack information throughout the integration to improve its capabilities while creating normal behavior profiles. NSL-KDD dataset was used for evaluating this research. According to the results, the proposed strategy outperformed existing methods in terms of detection rate for both unknown and known attacks while retaining a low false-positive rate. In addition, the proposed solution considerably decreases the high time complexity of the training and testing processes.

Table - 1: Summary Table of the Reviewed Papers

Paper	Authors	Year	Dataset	Algorithm Used	Accuracy (%)
Machine Learning Based Intrusion Detection System	Anish and Sundarakantham[3]	2019	NSL-KDD	SVM and Naïve Bayes	SVM- 93.95 Naïve Bayes- 71.001
A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks	Chuanlong et al.[11]	2017	NSL-KDD	RNN	97.09
Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms	Jiadong et al. [15]	2019	UNSW-NB15	Random Forest	92.8
A Machine Learning Approach for Intrusion Detection System on	Iram et al. [2]	2020	NSL-KDD	SVM, RF, ETC, DT, KNN, MLP, LR and	Above 99 on RF, ETC, and DT

NSLKDD Dataset				NB	
Enhanced Network Anomaly Detection Based on Deep Neural Networks	Naseer et al. [14]	2015	NSL-KDD	ELM, KNN, DT, RF, SVM, NB, QDA	DCNN and LSTM models showed exceptional performance with 85% and 89% Accuracy
Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks	Abdulsalam et al. [10]	2021	NSL-KDD	XGBoost	95.55
Artificial Intelligence based Network Intrusion Detection with Hyper-Parameter Optimization Tuning on the Realistic Cyber Dataset CSE-CICIDS2018 using Cloud Computing	Kanimozhi and Jacob [16]	2019	CSE-CIC-IDS2018	ANN	99.97

V. CONCLUSION

An intrusion detection system (IDS) is a powerful technology in the network security domain. Machine Learning techniques have been widely used in network security and have served as a handy tool in various network security disciplines. However, it is unrealistic to expect existing security methods to prevent network intrusion. As a result, a machine learning-based intrusion detection system is a critical component of cybersecurity. Furthermore, NIDS offers many potential advantages over the traditional system, such as lowering the number of human resources required for monitoring, improving detection accuracy, and assisting the information security community in learning about emerging threats. Based on the findings, it can be concluded that machine learning has much promise in Network Intrusion Detection when used effectively.

This survey has explored many published papers in IDS using machine learning algorithms. The survey also evaluates different classification techniques based on machine learning for the NIDS. This paper explores the ML algorithms used for IDS in other datasets and compares the results. This study will demonstrate that the algorithm and application area affect the detection rate, false-positive rate, and accuracy. In the future, extensive research will be conducted on ML algorithms to provide a better solution for the IDS using CSE-CIC-IDS2018 on AWS dataset.

VI. REFERENCES

[1] Steve Morgan, "Cybercrime To Cost The World $10.5 Trillion Annually By 2025,"Cybercrime Magazine, Nov. 10, 2020. https://cybersecurityventures.com/cybercrime-damage-costs-10-trillion-by-2025/ (accessed Feb. 03, 2022).

[2] I. Abrar, Z. Ayub, F. Masoodi, and A. M. Bamhdi, "A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset," in 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, Sep. 2020, pp. 919–924. doi: 10.1109/ICOSEC49089.2020.9215232.

[3] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, Apr. 2019, pp. 916–920. doi: 10.1109/ICOEI.2019.8862784.

[4] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, Jul. 2009, pp. 1–6. doi: 10.1109/ CISDA.2009. 5356528.

[5] J. L. Leevy and T. M. Khoshgoftaar, "A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data,"J Big Data, vol. 7, no. 1, p. 104, Dec. 2020, doi: 10.1186/s40537-020-00382-x.

[6] "IDS 2018 | Datasets | Research | Canadian Institute for Cybersecurity | UNB." https://www.unb.ca/cic/datasets/ids-2018.html (accessed Feb. 03, 2022).

[7] R. Gandhi, "Naive Bayes Classifier,"Medium, May 17, 2018.https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed Feb. 03, 2022).

[8] N. Naidu and Dr. R. V. Dharaskar, "An Effective Approach to Network Intrusion Detection System using Genetic Algorithm,"IJCA, vol. 1, no. 3, pp. 26–32, Feb. 2010, doi: 10.5120/89-188.

[9] Nikhitha M. and Jabbar M.A., "K Nearest Neighbor Based Model for Intrusion Detection System,"ijrte, vol.

8, no. 2, pp. 2258–2262, Jul. 2019, doi: 10.35940/ijrte.B2458.078219.

[10] A. O. Alzahrani and M. J. F. Alenazi, "Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks,"Future Internet, vol. 13, no. 5, p. 111, Apr. 2021, doi: 10.3390/fi13050111.

[11] C. Yin, Y. Zhu, J. Fei, and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,"IEEE Access, vol. 5, pp. 21954–21961, 2017, doi: 10.1109/ACCESS.2017.2762418.

[12] M. Monshizadeh, V. Khatri, B. G. Atli, R. Kantola, and Z. Yan, "Performance Evaluation of a Combined Anomaly Detection Platform,"IEEE Access, vol. 7, pp. 100964–100978, 2019, doi: 10.1109/ ACCESS.2019. 2930832.

[13] Z. K. Ibrahim and M. Y. Thanon, "Performance Comparison of Intrusion Detection System Using Three Different Machine Learning Algorithms," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, Jan. 2021, pp. 1116–1124. doi: 10.1109/ ICICT50816.2021. 9358775.

[14] 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.

[15] J. Ren, J. Guo, W. Qian, H. Yuan, X. Hao, and H. Jingjing, "Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms,"Security and Communication Networks, vol. 2019, pp. 1–11, Jun. 2019, doi: 10.1155/2019/7130868.

[16] V. Kanimozhi and T. P. Jacob, "Artificial Intelligence based Network Intrusion Detection with Hyper-Parameter Optimization Tuning on the Realistic Cyber Dataset CSE-CIC-IDS2018 using Cloud Computing," in 2019 International Conference on Communication and Signal Processing (ICCSP), Apr. 2019, pp. 0033–0036. doi: 10.1109/ICCSP.2019.8698029.

[17] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection,"Expert Systems with Applications, vol. 41, no. 4, pp. 1690–1700, Mar. 2014, doi: 10.1016/j.eswa.2013.08.066.