International Journal of Science Academic Research
Vol. 05, Issue 01, pp.6915-6922, January, 2024
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

LITERATURE REVIEW OF IMPLEMENTATION OF MACHINE LEARNING ALGORITHMS FOR IMPROVING THE NETWORK SECURITY

***Ayantika Sarkar and Pradnya Kashikar**

Department of CSIS, Birla Institute of Technology and Science, Pilani, India

Received 20th November 2023; **Accepted** 16th December 2023; **Published online** 30th January 2024

Abstract

Network security is becoming a top priority for people, enterprises, and governments as the digital world advances. Innovative and flexible solutions will be needed due to the growing complexity and diversity of cyberthreats. Machine learning (ML) has become an effective means for improving network security because it can quickly identify, cease, and neutralize many kinds of threats. There are several uses for machine learning in the realm of network security. The applications of machine learning in network security are divided into two categories in this paper: Malware detection system and Intrusion Detection System (IDS) – Signature based IDS and Anomalybased IDS. A few machine learning techniques, such as Supervised learning, Unsupervised learning, and Reinforcement learning, that have been utilized in the field of network security and the threat landscape for network security between 2020 and 2023 is also discussed in this paper. Finally, a literature review of the machine learning techniques in the field of network security have also been discussed based on the survey of various research works.

Keywords: Anomaly, Malware Detection, Unsupervised Learning, Reinforcement Learning, Supervised Learning, Intrusion Detection.

1. INTRODUCTION

One of the main factors for the continuous transition to a networked, information society is the uninterrupted and secure functioning of high performance communication networks. In a similar vein, important network infrastructure continues to be a prime target for impersonation attacks that compromise communication availability, confidentiality, or integrity. Current network monitoring systems provide high dimensional network data, which makes it possible to use machine learning techniques widely to enhance the identification and categorization of aberrant occurrences. As we are advancing towards an era of high network usage, the protection of our networks from various intrusions and malware attacks should be one of our priorities. The analysis of current threat landscape reports show that cyber attacks has tripled from 2018 to 2023. An aspect of artificial intelligence known as machine learning has become a potent weapon in the continuing fight to protect data and digital assets. Real time threat detection, mitigation, and response to cyber attacks may be achieved using machine learning in cyber security. By analyzing large and complicated information, finding abnormalities, and making predictions, machine learning algorithms open new possibilities for improving cyber security solutions. In this paper, we give an overview of the network threat landscape from 2020-2023, followed by the various machine learning (ML) techniques which have been used in the field of network security. The machine learning (ML) techniques discussed include Supervised Learning, Unsupervised Learning and Reinforcement Learning. Different categories of each of these techniques are discussed. Also, a review of the exsting machine learning (ML) based algorithms by analysis of multiple research projects.

Corresponding Author:** ***Ayantika Sarkar
Department of CSIS, Birla Institute of Technology and Science, Pilani, India.

2. Objectives and Scope

The objectives and scope of this paper are as follows:

- To give a summary of the threats facing networks between 2020 and 2023.
- To describe and explore the machine learning algorithms that are most frequently employed in network security.
- To assess each algorithm's advantages and disadvantages in relation to applications related to network security.
- To analyze and contrast the various machine learning algorithms' performance indicators in relation to network security.
- To draw attention to the metrics, such as detection accuracy, false positive/negative rates, and scalability, that are used to evaluate the efficacy of these algorithms.

3. Network Threat Landscape

According to [1] there has been an increase in phishing attacks by 81% as reported by global enterprises since the beginning of 2020. [10] reports $26.2 billion of losses in 2019 with Business E-mail Compromise (BEC) attacks. The findings from [2] indicated that the total amount of money lost to identity theft in 2020 was $56 billion, with $13 billion coming from classic identity fraud and $43 billion from identity fraud schemes. In [10], it has been stated that during the one month period (end of February 2020 to end of March 2020), the number of phishing assaults containing the virus climbed by 667%. Report [3] revealed the global average cost of data breach to be $4.24m in 2021. The report [4] shows that in December 2021, a record of over 316,000 phishing attacks is reached globally. The analysis in [28] mentions that over 2021, there has been a greater emphasis on business models including "Ransom ware as a Service" (RaaS), which has made it challenging to properly attribute specific threat actors. Over the course of 2021, triple extortion ransom ware tactics become more common. Malware developers keep coming up with new techniques to impede dynamic analysis and reverse engineering. In comparison to the previous several years, the number of crypto jacking infections reached a record high in the first quarter of 2021. Threat actors were motivated to carry out these assaults by the financial benefit connected with crypto jacking. In the year of 2021, [9] revealed that there is a record amount of crypto currency mining and crypto jacking activity. The analysis done in [8] shows that in 2022, phishing has become more common, sophisticated, and context based in approach of the attack with the increase in numbers. Consent phishing is one of the types of phishing that is taking a rise in count. Ransom ware attacks are becoming more complex along with moving towards IoT and mobile networks. Attacks by cybercriminals are becoming more frequent and affect vital infrastructure. The two most popular ways to get infected with ransom ware are still through phishing emails and brute forcing Remote Desktop Services (RDP) connections. According to [5] networked device counts will rise from 18.4 billion in 2018 to 29.3 billion by end of 2023. A broad range of Internet of Things (IoT) applications will be supported by around half of those connections (14.7 billion by 2023 compared to 6.1 billion in 2018). It also shows an estimation that the distributed denial of service (DDoS) attacks will rise to 15.4 million in 2023 which is almost double of 7.09 million which were the number of DDoS attacks in 2018. Fig. 1 shows the increase in DDoS attacks from 2018-2023.

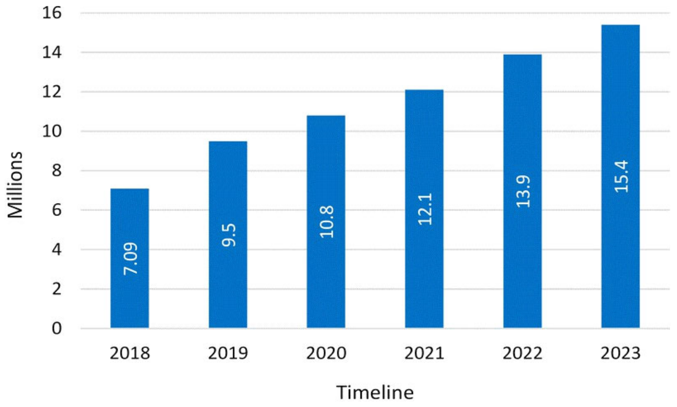

Fig. 1. Increase in Ransom ware attacks from 2018-2023 [5]

4. Machine Learning Techniques In Network Security

4.1. Supervised Learning

Supervised learning, a machine learning paradigm, uses a collection of paired input output training samples to determine the input output connection information of a system. The input output training sample is also known as labeled training data or supervised data as the output is thought of as the label of the input data or the supervision. Machine learning models constantly learn in the field of network security by examining data to identify patterns that improve malware detection in encrypted traffic, identify insider threats, identify online intruders to keep users safe, or safeguard cloud data by identifying questionable user activity. Algorithms of supervised learning can be categorized into two (2) parts, namely -

4.1.1. Classification

An algorithm is used in classification to precisely place test data into designated groups. It identifies entities in the dataset and tries to make recommendations on the definition or labeling of those items. Some of the classification algorithm include:

- **Support Vector Machine (SVM)**: A common supervised learning model for data regression and classification is the support vector machine. Nevertheless, it is usually applied to classification difficulties, creating a hyperplane where the maximum distance exists between two classes of data points. The decision boundary is a hyperplane that divides the classes of data points on either side of the plane. Fig. 2 shows the architectural diagram of the Support Vector Machine (SVM) which consists of an input layer, a hidden layer, and an output layer where the bias is applied. The bias is applied at the last step, post which the output of the classification is produced.

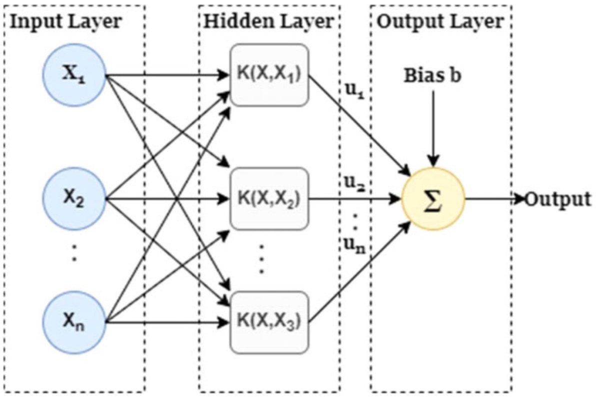

Fig. 2. Architecture of Support Vector Machine (SVM) [34]

- **K-Nearest Neighbor (KNN)** : The KNN algorithm, which is often referred to as K-nearest neighbor, is a non-parametric technique that groups data points according to their proximity and correlation with other accessible data. It is assumed by this technique that comparable data points can be located close to one another. Consequently, it aims to determine the separation between data points, typically using the Euclidean distance, and then designates a category according to the most prevalent category or mean.

- **Random Forest**: Another adaptable supervised machine learning approach for both regression and classification is called random forest. The term "forest" alludes to a grouping of uncorrelated decision trees that are combined to lower variance and produce more precise data predictions. Fig. 3 shows the architecture of Random Forest algorithm. The input data X is divided into "b" number of trees and output from each of the tree is expressed as k_1, k_2,. ,k_b. The final output k is expressed by averaging or classification, depending on the type of input data X.

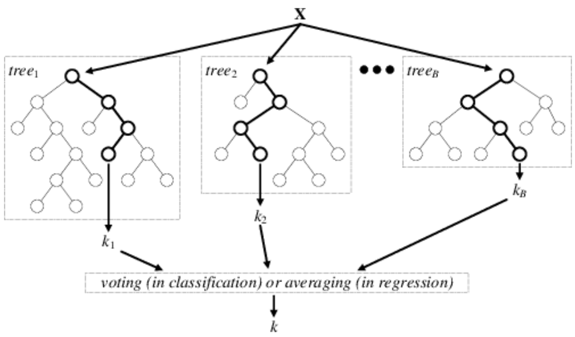

Fig. 3 Architecture of Random Forest [35]

4.1.2. Regression

To comprehend the link between dependent and independent variables, regression is utilized. Few regression algorithms are:

1. **Linear Regression** : The link between a dependent variable and one or more independent variables can be found using linear regression, which is then commonly used to forecast future events. When there is just one independent variable and one dependent variable, simple linear regression is utilized. It is known as multiple linear regression because the number of independent variables increases. The goal of each kind of linear regression is to depict the line of best fit, which is determined using the least squares approach. On a graph, nevertheless, this line is straight, in contrast to other regression models. Fig. 4 shows the architecture of linear regression algorithm, where x_1, x_2, x_3 and x_4 refers to the input, b is the bias, θ_1, θ_2, θ_3 and θ_4 are the parameters or the weights of the model and y is the output of the model.

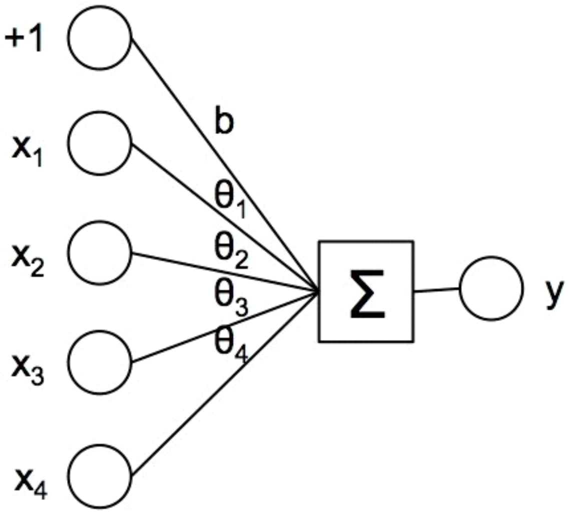

Fig. 4 Architecture of Linear Regression [36]

- **Logistic Regression** : Logistic regression is used when the dependent variable is categorical, i.e., it has binary outputs, like “yes” or “no”, whereas linear regression is utilized when the dependent variables are continuous. While the goal of both regression models is to comprehend the connections between data inputs, the primary use of logistic regression is in the resolution of binary classification issues, including spam detection. Fig. 5 shows the architecture of logistic regression.

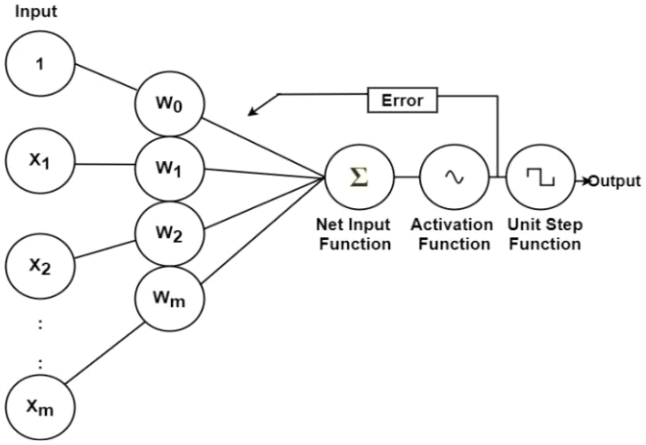

Fig. 5 Architecture of Logistic Regression [37]

There are multiple applications of the supervised learning algorithms, primarily, KNN, Random Forest, SVM in the field of anomaly detection.

4.2. Unsupervised Learning

Unsupervised learning analyzes and groups unlabeled datasets and finds hidden relationships or patterns in the data without requiring human interference. Unsupervised learning can be categorized into three (3) kinds-

4.2.1. Clustering

Using the clustering approach, unlabeled data may be grouped according to their similarities or differences. Algorithms for clustering are used to process unclassified, raw data items into groups that are represented by informational structures or patterns. There are several types of clustering algorithms, including the following:

2. ***Hierarchical Clustering***: Hierarchical cluster analysis (HCA) is another name for this type of unsupervised clustering technique. It is considered that agglomerative clustering is a “bottoms up approach”. Its data points are first separated into several groups before repeatedly being combined based on similarity until a single cluster is formed. Two types of hierarchical clustering algorithms are as follows:

1. **Agglomerative** : The following four approaches are frequently used to calculate similarity:
   - *Ward’s connection* : According to this technique, the increase in the sum of squared after the clusters are combined defines the distance between two clusters.
   - *Average linkage*: The mean distance between two sites in each cluster defines the average linkage method.
   - *Complete (or maximum) linkage* : The maximum distance between two locations in each cluster defines this technique.
   - *Single (or minimum) linkage* : The smallest distance that separates two locations in each cluster defines this technique.

2. **Divisive**: Divisive clustering is the antithesis of agglomerative clustering, utilizing a “top down” methodology. In this instance, the distinctions between data points are used to split a single data cluster. Even though it is not frequently employed, hierarchical clustering nevertheless makes divisional clustering relevant. A dendrogram, a figure resembling a tree that shows how data points merge or divide at each iteration, is typically used to display these clustering processes.

- *Probabilistic clustering* : An unsupervised method for resolving density estimates or “soft” clustering issues is a probabilistic model. Data points are grouped using probabilistic clustering according to how likely it is that they will fall into a specific distribution. One of the most used probabilistic clustering techniques is the Gaussian Mixture Model (GMM).
- *Exclusive clustering* : A type of grouping known as exclusive clustering limits a data point to existence inside a single cluster. Another name for this is “hard” clustering. Example of exclusive clustering is K-Means

clustering, where, the data points are divided into K groups, where K is the number of clusters that may be found depending on how far one group's centroid is from the other. A specific centroid's nearest neighboring data points will be grouped together under that centroid's category, whereas, a lesser K number will indicate bigger groups and less granularity, a greater K value will be suggestive of smaller groupings with more granularity.

- *Overlapping clustering* : It is like exclusive clustering but allows data points to belong to multiple clusters with separate degrees of membership. For example – Fuzzy k-means clustering.

4.2.2 Association

A rule based technique for determining correlations between variables in each dataset is called an association rule. The Apriori method is the most often utilized of the several algorithms available for generating association rules, including Eclat, FP-Growth, and Apriori.

4.2.3. Dimensionality Reduction

Large datasets may be combed through by unsupervised learning models, which can then identify anomalous data pieces. These abnormalities may draw attention to defective machinery, mistakes made by people, or security lapses. Unsupervised learning is the perfect option in cybersecurity, where the attacker is constantly altering their tactics, because of its capacity to identify patterns and discrepancies in information. It is far superior in a scenario where the attacker is constantly changing forms since it does not search for a specific label; instead, any pattern that deviates from the standard will be marked as risky.

4.3. Reinforcement Learning

Because it can learn by itself by exploring and taking advantage of the unfamiliar environment, reinforcement learning (RL), a subfield of machine learning, is the closest kind of learning to human learning. RL is very flexible and helpful in real time and hostile contexts because it can simulate an autonomous agent to operate in an optimum manner in a sequential manner with or without prior knowledge of the environment. Fig. 6 shows the architecture of reinforcement learning.

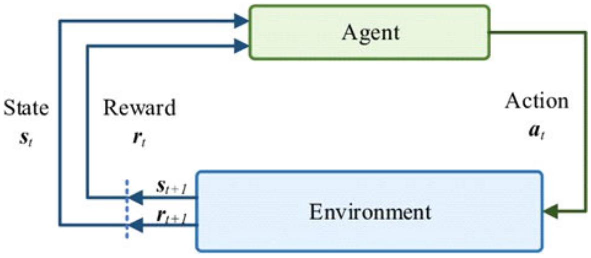

Fig. 6 Architecture of Reinforcement Learning [38]

Deep learning is sometimes integrated into reinforcement learning (RL) techniques, utilizing the capabilities of representation learning and function approximation to tackle several intricate challenges. Therefore, the deep learning and reinforcement learning a combination shows good fit for network security applications as cyberattacks are becoming more complex, fast, and common.

5. Machine Learning Applications In Network Security

With the onset of COVID-19, a drastic rise in the network threats has been observed as there is a global increase of usage of networking devices in all the sectors. To protect from the cyberattacks in the network, there are various intrusion detection methods which are implemented. As time progresses, the attacks are becoming much more sophisticated and to handle those attacks, machine learning (ML) is one of the methods which can be implemented.

5.1. Malware Detection

The report [27] says that after the 1990s, malware has seen significant modification. Initially, it was mostly made up of trivial programs created by programmers to demonstrate Windows vulnerabilities they had found. However, the focus of the current malware creation is on identity theft, fraud, and distributed denial of service (DDoS) assaults, payload attacks, etc. which makes it a far more significant issue for service providers and their customers. The research done in [29] shows that many malware programs connect with the attack originator via the Internet to get fresh assignments, software upgrades, or to release data they have gathered. Yet, such malware most likely employs a widely used network protocol to get past firewalls when it tries to interact with its Command and Control (C&C) center. Popular websites may occasionally be used as proxies or as part of the communication mechanism for malicious activity with the C&C center.

Malware programs can conceal themselves within systems or stop functioning when they detect attempts to identify them. As a result, it is necessary to employ passive systems (also known as trusted monitoring) that can identify malicious activity on targeted computers without requiring physical access. [23] introduces a behavioral malware clustering method at the network level that concentrates on HTTP based malware. Malware samples are grouped according to the idea of structural similarity between the malicious HTTP traffic they produce. The behavioral clustering technique reveals commonalities between malware samples through network level analysis that may be missed by existing system level behavioral clustering technologies. To categorize harmful and benign traffic and assign malicious activities to a malware family for both known and unknown malware, [23] provided a network classification approach. It also demonstrates how improved classification performance is achieved by varying the observation resolution, cross layers, and protocol characteristics. In addition to referring to several observation resolutions (transaction, session, flow, and conversation windows), the suggested model collects 972 behavioral elements from various protocols and network layers and reduces the data dimensionality to a manageable level. On this filtered data, while the very simple Naïve Bayes algorithm fails for some complex cases (e.g., APT1, Xpaj), it is effective enough in most cases. The Random Forest algorithm, which improves the J48 algorithm, can detect all new families with very high accuracy (most of the families detected with an area under the curve (AUC) of 0.98 except Conficker that detected with an AUC of 0.77). A well liked method for malware detection and malware family categorization is network behavioral modeling as discussed in [24]. Most of the research that has already been done focuses on certain malware kinds, such botnets, or on a particular kind of assault, like denial of service attacks, or anomaly detection in particular protocols or network layers, for example in the papers [25] and [26]. [33] suggests using distributed reinforcement learning to identify DDoS assaults that cause flooding. A sender agent (lower hierarchical level) learns semantic less communication signals that serve as a summary of its local state observations. Agents are arranged hierarchically. It is also necessary for the higher hierarchical level receiving agent to acquire the ability to decipher these signals lacking in meaning. [30] proposes a model which handles the SQL Injection from the client side instead of server. For feature extraction, Word Level Term Frequency and Inverse Document Frequency (TF- IDF) is used. The extracted data is executed using Machine Learning (ML) and Deep Learning (DL) algorithms – Convolutional Neural Network (CNN), Support Vector Machine (SVM), Passive Aggressive, Logistic Regression and Naïve Bayes. Out of all the algorithms CNN achieves the highest accuracy of 97%. The accuracies of SVM, Passive Aggressive, Logistic Regression and Naïve Bayes are 79%, 79%, 92% and 95% respectively. Fig. 7 shows the comparison of the accuracy of the algorithms. In [31], a hybrid model is proposed for the detection of SQLi attack by using Artificial Neural Network (ANN) and Support Vector Machine (SVM) which attains a precision of 99.54% and f1 score of 99.57%.

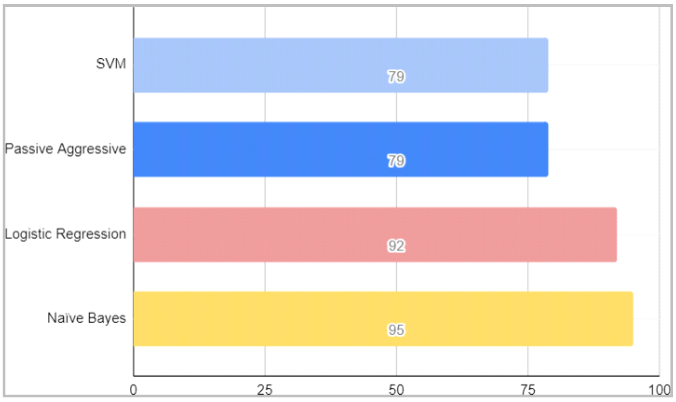

Fig. 7. Comparison of the accuracy in detection of malware

5.2. Intrusion Detection

In general, machine learning systems for intrusion detection can be implemented as numerous classifiers, which may or may not consider every feature or a feature subset of the used datasets, or as single classifiers (standalone units). Furthermore, depending on the chosen categorization method, these approaches can be categorized as either anomaly based or signature based. As a result, the existing survey research activities are categorized according to the machine learning techniques that have been used either signature or anomaly based module.

5.2.1. Signature based IDS

[11] shows the various approaches of signature based IDS along with its significance and the methods used to implement them. The paper describes about 12 types of approaches, namely – Network Behavior based IDS Approach, Knowledge based IDS Approach, Hierarchically Structured IDS Approach, Survey based IDS Approach, Virtual Switch based IDS Approach, Clustering based Approach, Feature Selection based Approach, Application based Approach, Data mining techniques based Approach, Classification based Approach, Expert System based Approach, Decision Tree based Approach. In [12], an ensemble machine learning strategy is constructed using six K-Nearest Neighbor (K-NN) and six Support Vector Machine (SVM) classifiers. The outcomes of all twelve models are combined using three different methods. The first scheme employs weights generated by PSO and integrates them using the Weight Majority Voting Algorithm (WMA); the second approach fine tunes the behavioral parameters of PSO using Local Unimodal Sampling (LUS); and the final scheme uses WMA to integrate the results produced by all classifiers. The three previously mentioned situations' performances are contrasted. Comparing the two approaches, LUS-PSO-WMA provides greater accuracy than the others. The accuracy for the Normal, Probe, DoS, U2R, and R2L types on the KDD'99 dataset is 83.6878%, 96.8576%, 98.8534%, 99.8029%, and 84.7615%, respectively, according to the reported findings. Fig. 8 shows the comparative study of the accuracy from the LUS-PSO-WMA approach. [32] develops a model based on Signature based IDS "Snort", Basic Analysis and Security Engine (BASE) and TCP Replay and is run on DARPA dataset. The model produces 42 unique alerts from the dataset. Since signature based IDS can be created using Network Behavior based IDS Approach, it can help in detection of malwares. Like [30] proposes a ML based model using Signature based detector for the detection of SQL Injection attacks.

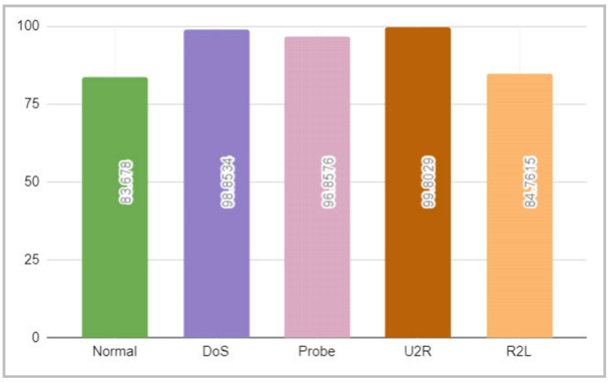

Fig. 8 Comparison of the accuracy from the LUS-PSO-WMA

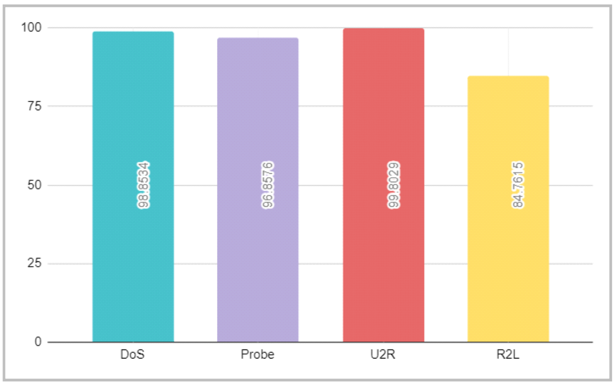

Fig. 9 Comparison of the detection rates on the various intrusions

The datasets are acquired through the testing of malfunctioning websites like bWapp and the Damn Vulnerability Web Application (DVWA). They executed the model on four (4) sets of test data in which the model achieved an accuracy of 93% on the first dataset and 100% on the fifth dataset. In [13] the purpose of an IDS framework with fuzzy association rules is to log the relationship between attack signatures and TCP/IP parameters. The use of high dimensional association rule mining is utilized, whereby the link between connected occurrences is clearly established by the identified rules. This approach expresses the logic (conjunction of numerous TCP/IP parameters) from the dataset, and the intrusion variations are shown by the findings. This method accurately and confidently detects attack signatures. On the 1998 DARPA dataset, detection rates of 99%, 95%, 75%, and 87% are attained for DoS, Probe, U2R, and R2L, respectively, using a six dimensional rule mining method. Fig. 9 shows the comparison of the detection rates of the various intrusions on the DARPA'98 dataset.

5.2.2. Anomaly based IDS

Based on support vector machines, [14] paper proposes the network intrusion detection system combining misuse and anomaly intrusion detection. In [15], an IDS framework is designed using the PSO-SVM technique on the KDD'99 dataset. Here, two feature selection techniques are applied: Binary PSO (BPSO), which selected 18 characteristics from the featured dataset out of 41, and information gain. For the DoS, Probe, R2L, and U2R types, the reported detection rate is 99.4%, 99.3%, 98.7%, and 98.5%, in that order. On the other hand, 84.2%, 89.4%, and 25% are stated as the accuracy for Probe, R2L, and U2R, respectively. [16] research proposes a hybrid anomaly based intrusion detection solution that relies on both Decision Tree and K-Nearest Neighbor (K-NN). To improve the performance of the suggested strategy, optimal data is extracted from the NSL-KDD dataset using a feature selection procedure. According to the experimental findings, the suggested strategy achieved a 99.6% accuracy rate, a 0.2% false alarm rate, and a positive detection rate of 99.7%.

On the KDD'99 dataset, the GA technique employs 10,000 occurrences in [17] for training and testing dataset samples. Eight qualities are selected using the PCA technique. It seems that even with a tiny population size of ten, only one rule, possibly from the final population, was enough to classify the data into two categories: normal and anomalous patterns. The performance parameters that are being evaluated here are accuracy and False Positive Rate (FPR), which are stated as follows: 10.8% and 93.49% for regular classes and 94.19% and 2.75% for attack patterns, respectively. [18] proposes a hybrid intrusion detection technique that depends on enhanced Fuzzy-C Means Clustering (FCM) and Support Vector Machine (SVM). Firstly, the technique groups the preprocessed training dataset using the Fuzzy-C Means Clustering (FCM) consolidating feature data gain ratio. This reduces the complexity of large scale datasets and enhances the SVM classifier's performance. To further identify the sort of assault, an SVM classifier is created for each group whose entropy exceeds a predetermined threshold. The hybrid intrusion detection technique may achieve 99.19% ac- curacy and 0.76% false alarm rate, according to the experi- ment conducted on the NSL-KDD dataset. This method works better in DoS, Probe, and R2L identification than other detection methods that also make use of the NSL-KDD dataset. Another anomaly based IDS strategy is used in [19], where Principal Component Analysis (PCA) is the main machine learning technique for categorizing attacks. For DoS, Probe, U2R, and R2L attack variations, the published findings for this study include detection rates of 99.2%, 80.7%, 88.5%, and 94.5%, and False Positive Rate (FPR) of 0.2, 4, 0.6, and 4, respectively, corresponding to each attack type. In [20] after being trained on the labeled KDD'99 dataset, ten machine learning algorithms were evaluated on unlabeled datasets. The speed and efficacy of these machine learning algorithms are examined by the researchers in relation to several chosen benchmarks, including the kappa statistic, accuracy by attack class, accuracy by root mean squared error, and the percentage of correctly classified instances of the classifier algorithms. They have done a benchmark comparison of the algorithms where they are ranked based on correctly classified instances. Random Forest attains an accuracy of 99.9794%, J48 Tree 99.9603%, Bagging 99.9524%, Support Vector Machine 99.9245%, Multilayer Perceptron 99.9245%, Bayes Net 99.667%, Radial Basis Function 99.3243%, AdaBoostM1 97.8576%, Naive Bayes 92.7794%, Stacking 56.8378%. Fig. 10 shows the comparison of the accuracies achieved by these algorithms.

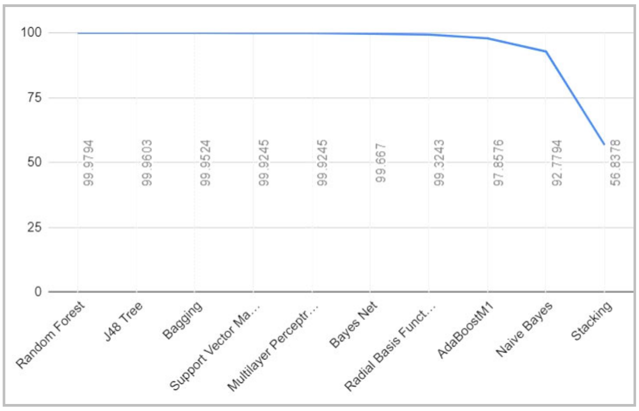

Fig. 10 Comparison of accuracy achieved by the algorithms

The [21] paper uses the Super Learner ensemble learning model to develop a novel detection method for network security and anomaly detection from [22]. A supervised learning technique called the Super Learner identifies the best mix of several foundational prediction algorithms. Utilizing five distinct ensemble combination algorithms for the Super Learner, two distinct scenarios, the well known MAWILab dataset for network attack detection, and a semi synthetic dataset for traffic anomaly detection in live cellular networks, the suggested solution is assessed. The suggested algorithm shows a detection accuracy of 92.8% for DDoS, 99.7% for mptp-la, 97% for netscan-ACK, 99.6% for netscan-UDP, 99.2% for ping flood.

6. Conclusion

As the usage of the network and automated resources are increasing, the cyberattacks and threats are evolving in their approaches and the techniques for cyberattacks are growing. Machine Learning (ML) has completely changed the way we think about security because of its capacity to go through enormous and complicated datasets, spot abnormalities, and adjust to new threats. Machine learning algorithms have demonstrated their effectiveness in differentiating between benign and dangerous activity, allowing proactive security systems in a variety of fields, including malware analysis and intrusion detection. More work needs to be carried in this field so that we can achieve high accuracy and with time we can protect and detect the vulnerabilities without the intervention of humans. This paper discusses about the various research works done in network security using machine learning (ML) algorithms. It also provides an overview of the threat landscape of network and gives brief description of the various machine learning algorithms that are implemented in the field of network security.

7. Declarations

7.1. Competing interests

- The authors have no competing interests to declare that are relevant to the content of this article.

7.2. Compliance with Ethical Standards

- This article does not contain any studies with human participants or animals performed by any of the authors.
- The authors did not receive support from any organization for the submitted work.

7.3. Research Data Policy and Data Availability Statements

- The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher.
- The manuscript contains third party material and obtained permissions are available on request by the Publisher.

REFERENCES

1. Ironscales (2023) The Ironscales State of Cyberse- curity Report: Blog, IRONSCALES. Available at: https:// ironscales.com/blog/ironscales-releases-findings- from-state-of-cybersecurity-survey/ (Accessed: 01 November 2023).
2. Buzzard, J. (2021) 2021 identity fraud study: Shifting angles, Javelin. Available at: https://javelinstrategy.com/2021-identity-fraud-study-shifting-angles (Accessed: 01 November 2023).
3. Data Endure | managed cybersecurity. it's about time. Available at: https://www.dataendure.com/wp-content/uploads/2021_Cost_of_a_Data_Breach_-2.pdf (2021) IBM Security (Accessed: 01 November 2023).
4. (2022) Quarter 2021 - APWG. Available at: https://docs.apwg.org/reports/apwg_trends_report_q4 _2021.pdf (Accessed: 03 November 2023).
5. Cisco annual internet Report - Cisco Annual Internet Report (2018–2023) White Paper (2022) Cisco. Available at: https://www.cisco.com/c/en/us/solutions/collateral/executive- perspectives/annual-internet-report/white-paper-c11-741490.html (Accessed: 01 November 2023).
6. Enisa threat landscape 2023 (2023) ENISA. Avail- able at: https://www.enisa.europa.eu/publications/enisa- threat-landscape-2023 (Accessed: 03 November 2023).
7. 2022 Internet Crime report - inter- net crime complaint center. Available at: https://www.ic3.gov/Media/PDF/AnnualReport/2022_ IC3Report.pdf (Accessed: 03 November 2023).
8. Enisa Threat Landscape 2022 (2023) ENISA. Avail- able at: https://www.enisa.europa.eu/publications/enisa- threat-landscape-2022 (Accessed: 04 November 2023).
9. Enisa Threat Landscape 2021 (2022) ENISA. Avail- able at: https://www.enisa.europa.eu/publications/enisa- threat-landscape-2021 (Accessed: 03 November 2023).
10. Enisa Threat Landscape 2020 - phish- ing (2021) ENISA. Available at: https://www.enisa.europa. eu/publications/phishing (Accessed: 04 November 2023).
11. Snehi, J. V. (2020, July). Diverse Methods for Signature based Intrusion Detection Schemes Adopted. Research Gate. https://www.researchgate.net/ publication/342691019_Diverse_Methods_for_Signature_ based_Intrusion_ Detection_Schemes_Adopted
12. Aburomman A. and M. B. I. Reaz, "A novel svm K NN pso ensemble method for intrusion detec- tion system," *Applied Soft Computing,* vol. 38,2016, pp. 360–372.]
13. Brahmi H., B. Imen, and B. Sadok, "OMC-IDS: At the cross roads of OLAP mining and intrusion detection," in Advances in Knowledge Discovery and Data Mining. New York, NY, USA: Springer, 2012, pp. 13–24
14. Xu T. and H. Hou, "Network intrusion detection based on support vector machine," Research Gate, https://www.researchgate.net/publication/261047632_ Network_Intrusion_Detection_Based_on_Support_ Vector_Machine (accessed Nov. 5, 2023).
15. Saxena H. and V. Richariya, "Intrusion Detection in KDD99 Dataset using SVM-PSO and Feature Reduction with Information Gain," *Int. J. Comput. Appl.* vol. 98, no. 6, 2014 pp. 25–29.
16. Foroushani Z.A., Y. Li, "Intrusion Detection System by Using Hybrid Algorithm of Data Mining Technique", 7th International conference on software and computer applications, 2018, pp. 119-123.
17. Khan, S. "Rule-based network intrusion detection using genetic algorithms," *Int. J. Comput. Appl.,* vol. 18, no. 8, 2011,pp. 26–29.
18. Zhang Z., P.Pan, "A Hybrid Intrusion Detection Method Based on Improved Fuzzy C-Means and Support Vector Machine", *International conference on communications, information system and computer engineering (CISCE),* 2019, pp. 210 -214.
19. Wang W. and R. Battiti, "Identifying intrusions in com- puter networks with principal component analysis,", First international conference on Availability, Reliability and Security(ARES06),2006, pp-279-286.
20. Ali A., Y.-H. Hu, C.C. (George) Hsieh, and M. Khan, "A comparative study on machine learning al- gorithms for network defense," ODU Digital Com- mons, https://digitalcommons.odu.edu/vjs/vol68/iss3/1/ (accessed Nov. 5, 2023).
21. Vanerio J. and P. Casas, "Ensemble learning approaches for network security and anomaly detection: Proceedings of the workshop on Big Data Analytics and machine learning for Data Communication Networks," ACM Conferences, https://dl.acm.org/doi/10.1145/3098593.3098594 (accessed Nov. 5, 2023).
22. Van der Laan M., E. C. Polley and A. E. Hubbard, "Super learner", in Statistical applications in genetics and molecular biology, vol. 6 (1), pp. 1-21, 2007.
23. Bekerman D., B. Shapira, L. Rokach, and A. Bar, "Unknown malware detection using network traffic classification," 2015 IEEE Conference on Communications and Network Security (CNS) (accessed Nov. 5, 2023).
24. Nari S. and A. A. Ghorbani, "Automated Malware Classification based on Network Behavior," in IEEE International Conference on Computing, Networking and Communications (ICNC), San Diego, CA, USA, 2013.
25. Sangkatsanee P., N.Wattanapongsakorn and C. Charnsripinyo, "Practical real time intrusion detection using machine learning approaches," *Computer Communications,* vol. 34, no. 18, pp.2227-2235, 1 December 2011
26. Stone Gross B, M. Cova, L. Cavallaro, B. Gilbert, M.Szydlowski, R. Kemmerer, C. Kruegel and G. Vigna,

"Your Botnet is My Botnet: Analysis of a Botnet Takeover," in 16th ACM Conference on Computer and Communications Security (CCS), Chicago, Illinois, USA, 2009.

27. "The case for network based malware detection tmcnet," Kindsight Security Labs, https://www.tmcnet.com/tmc/whitepapers/documents /whitepapers/2014/9599-case-network-based-malware- detection.pdf (accessed Nov. 5, 2023).
28. "The threat landscape in 2021," Symantec Enterprise Blogs, https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/threat-landscape-2021 (accessed Nov. 6, 2023).
29. Perdisci R., W. Lee, and N. Feamster, "Behavioral Clustering of HTTP Based Malware and Signature Generation Using Malicious Network Traces," Research Gate, https://www.researchgate.net/publication/220831984_Behavioral_Clustering_of_HTTP-Based_Malware_and_ Signature_Generation_Using_Malicious_Network_Traces (accessed Nov. 6, 2023).
30. Azman M. A., M. F. Marhusin, and R. Sulaiman, "Machine learning based technique to detect SQL injection attack," *Journal of Computer Science*, vol. 17, no. 3, pp. 296–303, 2021. doi:10.3844/jcssp.2021.296.303
31. Demilie W. B. and F. G. Deriba, "Detection and prevention of SQLI attacks and developing compressive framework using machine learning and hybrid techniques," *Journal of Big Data,* vol. 9, no. 1, 2022. doi:10.1186/s40537-022-00678-0
32. Kumar V. and O. P. Sangwan, "Signature Based In trusion Detection System Using SNORT," *International Journal of Computer Applications & Information Technology,* vol. I, no. III, Nov. 2012.
33. Servin, A. & Kudenko, D. (2008b). Multiagent reinforcement learning for intrusion detection: A case study and evaluation. In R. Bergmann, G. Lindemann, S. Kirn, & M. Pchouek (Eds.), Multiagent System Technologies, volume 5244 of Lecture Notes in Computer Science (pp. 159–170). Springer Berlin Heidelberg.
34. Vaibhav K., Kumar V C, Satish, B., Arunkumar, W., Seema, K., Pooja & P., Shruti. (2021). Enhancing Surface Fault Detection Using Machine Learning for 3D Printed Products Applied System Innovation. 10.3390/asi4020034.
35. Antanas, V.,V., Evaldas, G., Adas, P., M James & O., M. Charlotte. (2016). Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and Prediction of Shot Effectiveness.10.3390/s16040592.
36. Jean Pierre, B., H., Gaëtan & P., Francois. (2017). Deep Learning Techniques for Music Generation - A Survey.
37. Sandipan, B., G., Shivnath, R., Sandip, B., Rajesh & S., Sanjay. (2023). A Study of Stock Market Prediction through Sentiment Analysis.10.12723/mjs.64.6.
38. Rentao, G., Y., Zeyuan & Ji, Yuefeng. (2020). Machine Learning for Intelligent Optical Networks: A Comprehensive Survey.

****