A Comparison Analysis of Machine Learning Algorithms for Intrusion Detection

[1]Nishi Patwa, [2]Prachi Shah, [3]Pooja Thakkar

[1,2,3]Assistant Professor
[1]Computer Engineering, [2,3]Information Technology
[1,2,3]U. V. Patel College of Engineering, Ganpat University, Mehsana, India.

***Abstract*—Intrusion detection is a critical component of network security, as it helps identify potential threats and attacks on the system. The paper begins by discussing the introduction of the intrusion detection system and different types of intrusion detection systems. It then provides a comprehensive review of various machine learning algorithms, including unsupervised as well as supervised techniques. The paper also highlights the merits and demerits of various algorithms and the factors that affect their performance in intrusion detection. Finally, the paper concludes with a comparison of ML techniques used for developing IDS and summaries of relevant research articles.**

***Index Terms*—IDS, Intrusion Detection, Machine Learning, Cybercrime**

I. Introduction

Cyber-attacks are becoming more and more common, and organizations need to be prepared for them. Intrusion Detection Systems (IDS) are an important tool in the fight against cybercrime [11]. Intrusion detection systems (IDS) play a vital role in keeping your data and assets safe from unauthorized access or malicious attacks. It is designed to detect any unauthorized access attempts by monitoring network traffic for malicious activities such as port scans, buffer overflows, and worm propagation. Additionally, the system can also be used to monitor internal systems for unusual behavior or changes in user behavior that may indicate an attack [6]. By analyzing the data, it collects, an IDS can detect suspicious behavior such as malware infections, brute-force attacks, and network scanning. By detecting these threats early on in the process, an IDS can help organizations protect their systems from potential damage or loss of data.

The purpose of an IDS is to scrutinize the network traffic and detects any suspicious activity, such as unauthorized access attempts, malicious code execution, and data manipulation. It can also detect known threats such as viruses, worms, Trojans, and other malware. In addition to these threats, an IDS can also detect potential threats such as denial of service attacks and buffer overflows.

Types of Intrusion Detection Systems

An IDS or intrusion detection system is a software application that monitors network or system activities for malicious or unauthorized behaviors and alerts administrators when such activities are detected. There are four main categories of intrusion detection systems: network intrusion detection systems, host-based intrusion detection systems, wireless intrusion detection systems, and database activity monitoring. The working scenario for how an intrusion detection system distinguishes between a legitimate request and a forged request is shown in Figure 1.

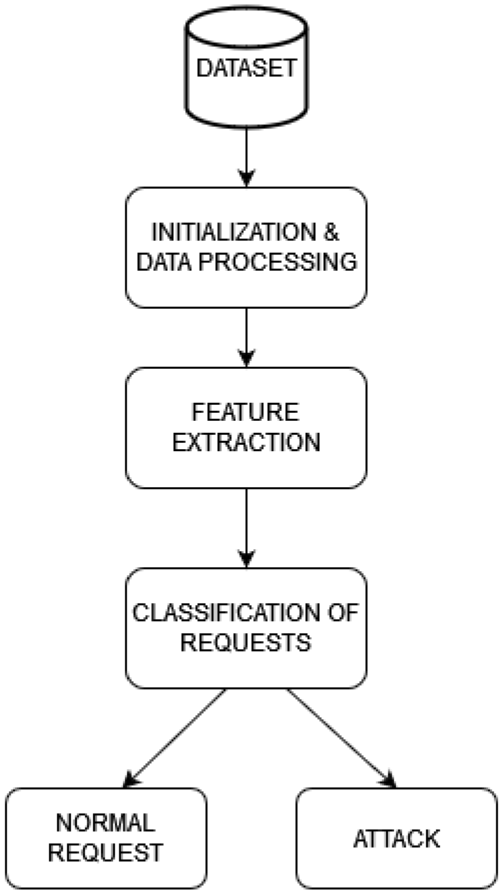

Figure 1. Flowchart for IDS

- Network Intrusion Detection Systems (NIDS): A NIDS is used to monitor incoming and outgoing network requests for spoofed activities. It is placed at a strategic point within a network. It analyses traffic patterns and compares them to known signatures of attacks. Some NIDS can also detect attempts to evade detection by looking for signs that an attacker is trying to cover their tracks.

- Host-Based Intrusion Detection Systems (HIDS): A HIDS resides on individual hosts or servers and monitors activity on that system for signs of an attack. It has intimate knowledge of what normal activity looks like on that particular system, making it more effective at detecting unusual or suspicious behavior. However, because it is installed on a single host, a HIDS cannot see activity occurring elsewhere on the network which may be part of an attack.

- Wireless Intrusion Detection Systems (WIDS): A WIDS monitors wireless traffic in order to detect unauthorized access points, rogue devices, and other Suspicious behavior. It can be used to supplement NIDS by providing visibility of traffic that would otherwise go undetected.

- Database Activity Monitoring (DAM): DAM tools monitor database activity in real-time, looking for SQL injection attacks, unauthorized access attempts, and other suspicious behavior. They can also provide insights into how databases are being used.

II. MACHINE LEARNING

Machine learning is changing the way we interact with technology. It’s no longer just about programming code to do something; it’s about using data to teach computers how to learn and make decisions. From self-driving cars to voice assistants, machine learning is becoming increasingly prevalent in our lives.

Machine learning is a subset of artificial intelligence. ML-based software learns from the experience gained through the training phase and accordingly makes predictions. If having a huge amount of data, ML Algorithms will improve the performance.

The primary goal of machine learning is to make computers learn on their own by increasing their ability to find patterns and insights in data. This is done through a process of training the computer with a dataset, which is then used to make predictions or recommendations.

Types of Machine Learning Techniques

Machine learning methodologies are split into two categories: supervised and unsupervised. Supervised learning methods are given a set of training data, which has both the input values (x) and the required output values (y). The algorithm learns from this data and is then able to generalize it to new data. They are used when we have a dataset with known labels [6]. The algorithm learns from the data and produces a model that can be utilized to predict the class of new data points. Unsupervised learning algorithms, on the other hand, only have input values (x) and must find structure in this data themselves in order to make predictions. Unsupervised learning methods are used in datasets with no labels. The algorithm tries to find patterns in the data and doesn't require any training data.

Pros of Machine Learning

Machine learning is a powerful tool that replicates human behavior where without any prior information it learns from experience and accordingly improves itself. It can be utilized to make predictions of upcoming events, identify patterns in data, and optimize decisions. Machine learning is already being used in a variety of fields such as finance, healthcare, marketing, and manufacturing.

Some benefits of machine learning include:

- Increased accuracy: Machine learning can provide accurate results than traditional methods because It can gain knowledge from training data and spot patterns that humans would miss.

- Automation: Machine learning algorithms can robotize all the things that previously needed human support, such as identifying spam emails or fraudulent credit card transactions.

- Faster results: Machine learning can often provide results faster than traditional methods because it can parallelize the work across multiple processors.

- Improved decision-making: It can be used to support industry to make accurate decisions by giving insights that would otherwise be unavailable. For example, machine learning could be used to help determine which products are most likely to sell well in a given market.

Cons of Machine Learning

ML is a powerful tool that can automatically extract patterns from data. However, machine learning is not perfect and has a number of limitations.

- One limitation of machine learning is that it can be biased if the data that is used to train the algorithm is itself biased. For example, if an algorithm is trained on data that is predominantly male, it may learn to associate male-associated traits with success and female-associated traits with failure [14]. This can lead to unfairness in decision-making when the algorithm is deployed in the real world.
- Another limitation of machine learning is its reliance on data. If the data used to train an algorithm is incomplete or inaccurate, then the algorithm will also be imperfect. This can lead to inaccurate predictions or decisions being made by the algorithm.
- Finally, machine learning algorithms are often opaque – they can be difficult for humans to understand how they arrive at their predictions or decisions. This lack of transparency can make it difficult to trust machine learning systems and could lead to important decisions being made without any accountability.

III. Machine Learning Classification Algorithms for Intrusion Detection

Support Vector Machine (SVM)

Supervised machine learning methods include Support Vector Machines. Both classification and regression can be done using SVM. The method may be trained using labeled data, and it can produce a hyperplane-based classification of the data into classes that optimize margin across all attack classes. SVM may conduct multi-class classification utilizing a cascade method in addition to being a binary classifier. SVM primarily depends on the parameters and types of kernels utilized [11]. The primary goal of SVM is to identify the ideal hyperplane that best divides the two groups. A relatively small number of support vector values are needed to define this hyperplane. The local minimum is not a problem for SVM, and it can handle noisy datasets. These characteristics fit the specifications needed to create an effective IDS [15].

SVM classifiers come in a variety of forms with varying functionalities [15]. SVMs can be divided into two primary categories: multiclass SVMs and binary SVMs, the latter of which can be broken further into linear and non-linear classes. Furthermore, the following categories can be applied to multi-class SVMs: One Against All (OAA) SVM, DAGSVM, and One Against One (OAO) SVM.

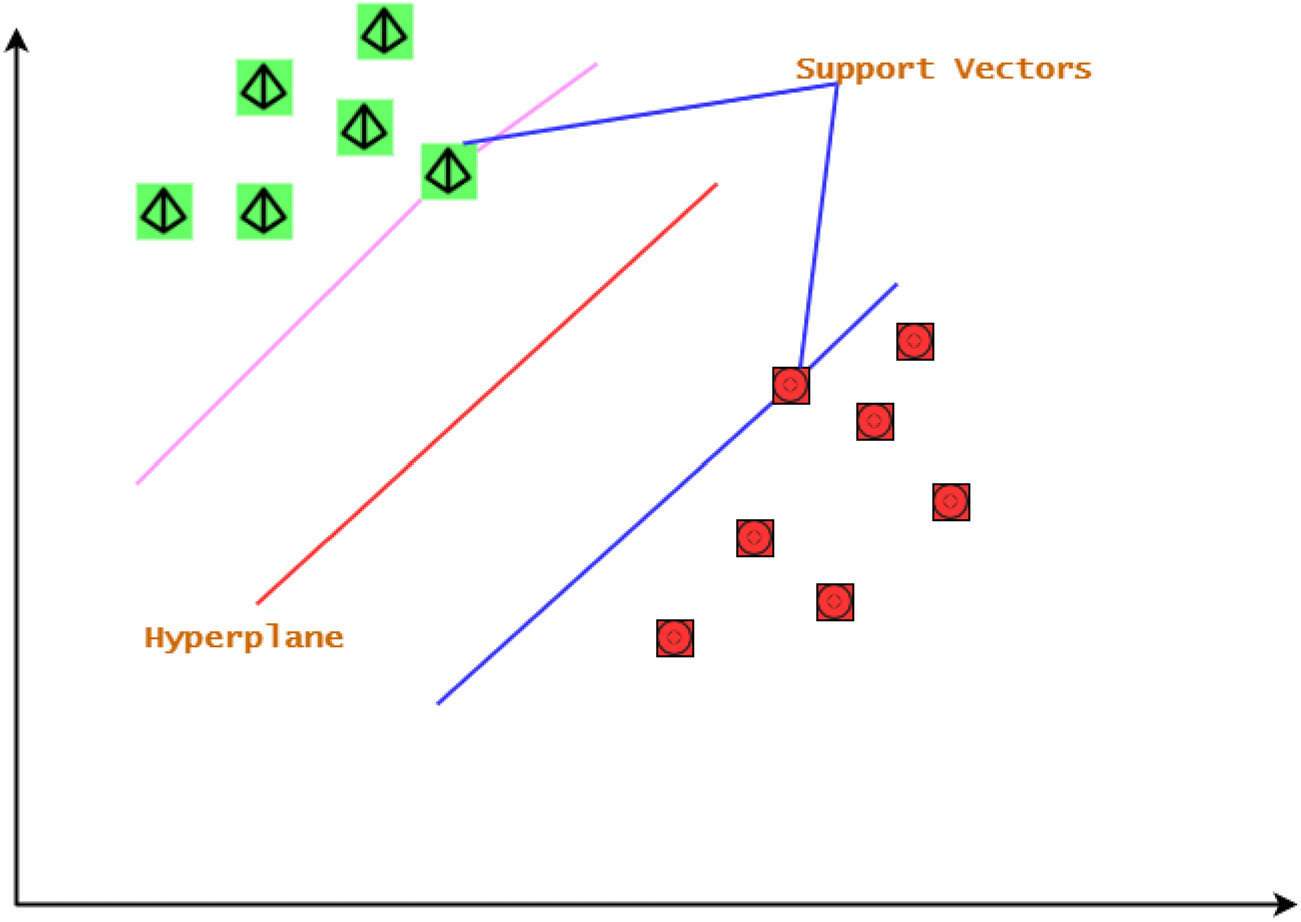

Figure 2. SVM Classifier

- **Categories of Binary SVM**

  There are two categories of binary SVM:

  a. Linear SVM : The data that can be linearly separated into two dimensions is identified as linearly separable data. In intrusion detection, we aim to separate the spoofed and the genuine ones. A linear SVM classifier is used to separate data into two unique classes through a straight line.

  b. Non-linear SVM : The data can be divided into more than two classes in any real-life scenario. Apart from the general scenario, the data can be divided into various attribute classes. In this kind of scenario, a non-linear SVM classifier can be applied. Non-linear SVM uses different kernels to classify data in multi-dimensions. Figure 2 illustrates how the SVM classifier chooses support vectors and selects a hyperplane.

- **SVM Algorithm Terminologies**

  a. Hyperplane: The SVM classifier generates a decision boundary for a given dataset. The SVM classifier algorithm first picks the best support vector for the given classification. The support vectors are then chosen as a base to formulate the decision boundary also known as Hyperplane [16]. The graphical representation of the support vector and hyperplane for 2-class classification in a two-dimensional space is shown in Figure 2. Hyperplanes can be obtained for binary as well as multiclass classification. The primary goal is to obtain a hyperplane with the highest margin.

  b. Support Vectors: Support vectors are the points that are nearest to the hyperplane. Based on the support vectors' decision of choosing a hyperplane. is taken by the algorithm.

K-Nearest-Neighbor

K-nearest neighbors algorithm belongs to the category of supervised machine learning. However, classification prediction problems are where it is most frequently applied [13]. Based on the training set provided, KNN projects the data points depending on the feature similarity. Every projected data point will be assigned a class based on the resemblance of the trained data points. In Fig. 3., we can see the classification of data points into various classes. KNN can be more powerful if the training data is huge. KNN is resilient to noisy training data [1]. Both supervised and unsupervised procedures can be carried out using the K-Nearest-Neighbour (KNN) ML technique. For instance, many of the clustering algorithms in use today are based on KNN.

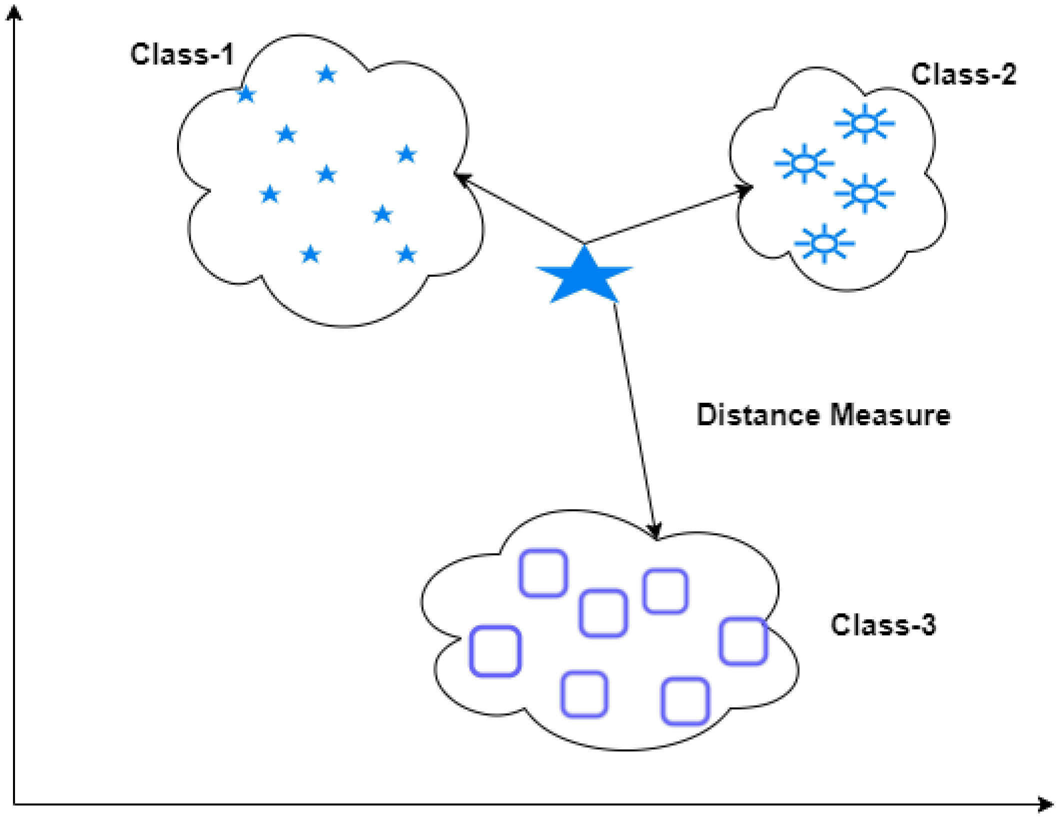

Figure 3. K-Nearest-neighbours

Naive Bayes

A straightforward probability classification known as naive Bayes is derived from Bayes' Theorem and assumes that each characteristic or variable is independent of every other [2]. Thomas Bayes, a British scholar, proposed Bayes' theorem as a method for forecasting possibilities based on past performance.

Decision Tree

Machine learning approaches are widely used to solve classification and regression problems. The decision tree is one such machine learning approach used to find the proper class of your data [8]. It is a tree-like structure with one parent node and subsequent child and leaf nodes. The classification decision is present on the leaf node.

Algorithm using decision trees:

1. First the process starts with the root node.
2. Second step is to find the best attribute for the next stage of classification. To find that attribute various techniques of attribute selection mechanisms are applied.
3. Create a decision tree with the best attribute in step 2.
4. Repeat the same process until the classification of the data is achieved.

As mentioned in step 2 we need to find attributes using the attribute selection mechanism. The three widely used mechanisms are:

- Gini Index: The Gini Index is a potent indicator of the randomness, impurity, or entropy in a dataset's values. Gini Index removes impurities from each level of the decision tree, starting from the top and step by step progressing to the bottom.

$$Gini = 1 - \sum_{i=1}^{n} (pi)^2$$

Where pi stands for the probability of classification.

- Entropy: Entropy is the measure o f purity. It measures how the feature is closely related to the available class. The below-mentioned formula shows how we can extract the entropy of our input.

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Where pi is the probability of our class and the + and - signs in the outcome show a positive and negative classification, also c describes the number of unique classes. If we have 10 data points and 3 points belong to the positive class and 7 points belong to the negative class the classification will have a negative sign.

- Information Gain: After the data cleaning the information gain methodology will extract the relevant and related information to our input. This information will provide details about how much our input is related to our classification. The gain value is used at each step to split a node.

$$Gain = E_{parent} - E_{children}$$

Random Forest

A random forest is a cluster of decision trees obtained from the input on different parameters. This technique uses hundreds or even thousands of Decision Trees to facilitate the learning or training of the algorithm. Each tree is built using just a few samples of the data (or just one sample). After all the decision trees are obtained a traditional voting system is executed to find the classification that received the highest vote. Decision tree lacks in giving accuracy whereas random forest gives us the upper hand as it has a voting system.

Artificial Neural Network

Artificial neural networks are developed to make computers behave how a human would behave in a particular situation. In essence, ANN is a replica of the human brain. ANN has the smallest functional element called the neuron, a working unit that behaves like a replica of the human brain. ANN comprises 3 layers: the first one is the Input layer, the successive or middle layer is called the Hidden layer, and the Output layer as shown in the figure. Like the response time of the human brain depends on the complexity of the situation to transfer that behavior to ANN different weights are applied to the neurons. ANN layers are shown in Figure 4. Neurons are activated in the hidden layer. With ANN a single perception model is developed.

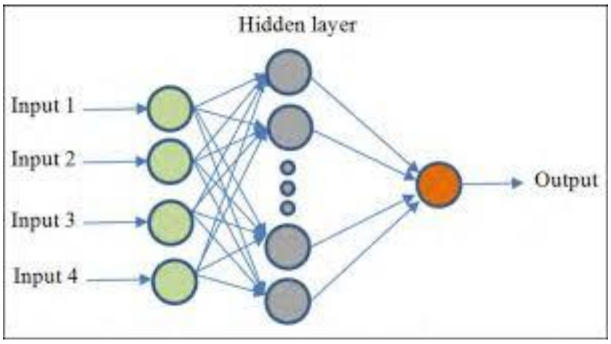

Figure 4. Artificial Neural Network

IV. COMPARISON OF DIFFERENT ML ALGORITHMS

The comparison study is done on the most popular data set, NSL-KDD serves as the benchmark for current internet traffic. The NSL-KDD data collection from Brunswick is an updated, improved version of the KDD'99. It has information of web traffic. This web traffic is recorded by a regular intrusion detection system. The dataset comprises of 43 features. Out of these 43 features, 41 are related to web traffic and remaining two are labels and scores. Labels indicate whether it is a normal request or attack and the score tells you the severity of the traffic input.

Table 1 Comparison of different ML Algorithms

TITLE	ALGORITHMS	ACCURACY	REMARK
Fast KNN Classifiers for Network Intrusion Detection System[1]	K-Nearest Neighbor	99.95%	High accuracy was attained. High computational time since feature selection cannot be applied.
Network Intrusion Detection using Supervised Machine Learning Technique with feature selection[2]	Artificial Neural Network	94.02%	Due to the use of feature selection, high accuracy was attained. Given the high false positive rate, the study cannot solve the issue of zero-day attacks.
Intrusion detection in computer networks using	Support Vector Machine and K-Means combined	96.81%	Integrating supervised and unsupervised learning

hybrid machine learning techniques[3]			methods increases the effectiveness of IDS.
Comparison of classification techniques applied for network intrusion detection and classification[4]	Random Forest Tree, Naïve Bayes Decision Tree	98.34% & 98.44%	Reduced the number of erroneous positives.
Random Forest Modeling for Network IDS[5]	Random forest	99.67%	The model is reliable because it has a high detection rate and a low rate of alerts. To increase accuracy, a feature selection technique like evolutionary computation must be used.
Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches[6]	K- Nearest Neighbor, KNN -Random Committee	98.727%, 99.696%	The accuracy of an ensemble technique is higher than that of a single classifier. Unable to tackle the problem of high-dimensional data.
Study on Decision Tree and KNN Algorithm for Intrusion Detection System[8]	Decision Tree, KNN	99.15%, 98.7%	Better results are produced by decision tree algorithms. Decision tree algorithm required significantly less time to develop than the KNN algorithm.
Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques[7]	Random Forest, Decision Tree	95.323%, 81.868%	High false positive rates and a poor detection rate.
Intrusion Detection Model Using Naive Bayes and Deep Learning Technique[10]	Naïve Bayes	99.93%	The suggested model's results show 99.9325 classification accuracy, 99.9738 detection rate, and 0.00093 false alarm rate.
A Feed-Forward ANN and Pattern Recognition ANN Model for Network Intrusion Detection[9]	Feed forward ANN, Pattern Recognition ANN	98.0792%, 96.6225%	Usage of considerable classifiers together enhances performance

V. CONCLUSION

The best algorithm for intrusion detection is one that uses machine learning because it is dependent on the problem's requirements and specific context. However, in general, k-Nearest Neighbors, Decision Trees, Random Forests, Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs) are some of the most widely used machine learning algorithms for intrusion detection. Decision Trees and Random Forests are often used for their simplicity and ability to handle categorical data. At the same time, SVMs are known for their ability to handle complex data and high-dimensional feature spaces. ANNs can be very effective for detecting complex patterns, and k-NN is often used when there is a need for fast detection. It is also important to consider the balance between accuracy and efficiency in real-time intrusion detection scenarios.

REFERENCES:

[1] Bobba Brao and Kailasam Swathi. (2017). Fast KNN Classifiers for Network Intrusion Detection System. Indian Journal of Science and Technology. 10(14). Researchgate. (1-10).

[2] Kazi A., Billal M. and Mahbubur R. (2019). Network Intrusion Detection using Supervised Machine Learning Technique with feature selection. International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). (pp. 643-646). IEEE.

[3] Deyban P. Miguel A. A, David P. A, and Eugenio S. (2017). Intrusion detection in computer networks using hybrid machine learning techniques. XLIII Latin American Computer Conference (CLEI). (pp.1-10). IEEE

[4] A.S.A. Aziz. (2016). Comparison of classification techniques applied for network intrusion detection and classification. Journal of Applied Logic 24. Elsevier, 109-118.

[5] Nabila Farnaaz and M.A Jabbar. (2016). Random Forest Modeling for Network Intrusion Detection System. International Multi-conference on information processing (IMCIP) 12 (pp. 213-217). Elsevier.

[6] Maniriho et al. (2020). Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches. International Journal of Intelligent Engineering and Systems. INASS. (433-445)

[7] Bhavani T. T, Kameswara M. R and Manohar A. R. (2020). Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques. International Conference on Sustainable Technologies for Computational Intelligence (ICSTCI). (pp. 637-643). Springer.

[8] Ashwini Pathak and Sakshi Pathak (2020). Study on Decision Tree and KNN Algorithm for Intrusion Detection System. International Journal of Engineering Research \& Technology (IJERT).

[9] Iqbal and Aftab. (2019). A Feed-Forward ANN and Pattern Recognition ANN Model for Network Intrusion Detection. International Journal of Computer Network and Information Security, 4. Researchgate (19-25)

[10] Mohammed Tabash, Mohamed Abd Allah, and Bella Tawfik (2020). Intrusion Detection Model Using Naive Bayes and Deep Learning Technique. The International Arab Journal of Information Technology.

[11] T.Saranyaa, S.Sridevi, C.Deisy, Tran Duc Chung, M.K.A.Ahamed Khan (2020). Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review. Third International Conference on Computing and Network Communications (CoCoNet'19).

[12] A. Bachar, N. E. Makhfi, O.E. Bannay (2020). Towards a behavioral network intrusion detection system based on the SVM model. 1st international conference on innovation research in applied science, engineering and technology (IRASET), Meknes, Morocco, pp. 1-7.

[13] Iram Abrar, Zahrah Ayub, and Faheem Masoodi, Alwi M Bamhdi (2020). A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC)

[14] Jaber, A.N., Rehman, S.U.(2020). FCM–SVM based intrusion detection system for cloud computing environment. Cluster Comput 23, 3221–3231.

[15] Mokhtar Mohammadi, Tarik A. Rashid, Sarkhel H.Taher Karim, Adil Hussain Mohammed Aldalwie, Quan Thanh Tho, Moazam Bidaki, Amir Masoud Rahmani, Mehdi Hosseinzadeh (2021). A comprehensive survey and taxonomy of the SVM-based intrusion detection systems. Journal of Network and Computer Applications, Volume 178.

[16] Ajdani, M, Ghaffary, H (2020). Design network intrusion detection system using support vector machine. Int J Commun Syst. 2021; 34:e4689