Intrusion Detection System Using Feature Selection and Machine Learning Techniques

Dr. Nandini Devi[1]*, Rashmi Saini[2], Balwinder Singh Dhaliwal[3] and Amit Deogar[4]

[1]*Electronics and Communication Engineering Department, National Institute of Technical Teachers Training and Research, Chandigarh, India*

[2]*Computer Science and Engineering Department, G. B. Pant Institute of Engineering and Technology, Pauri-Garhwal, Uttarakhand, India*

[3]*Electronics and Communication Engineering Department, National Institute of Technical Teachers Training and Research, Chandigarh, India*

[4]*Computer Science and Engineering Department, National Institute of Technical Teachers Training and Research, Chandigarh, India*

Corresponding: nandini.ece20@nitttrchd.ac.in

ABSTRACT

Intrusion is a serious problem in computer network security. With the rapid rise in various applications in the networking domain, intrusion attacks on networks increase that cannot be detected by humans effectively. To prevent such threats a system called Intrusion Detection System is designed. In this paper, five Machine Learning techniques i.e., K-Nearest Neighbours, Multiple Layer Perceptron, Decision Tree, Naïve Bayes, and Random Forest classifiers are used for Intrusion Detection System. For the proposed work, the NSL-KDD dataset with Random Forest Feature Selection Technique has been used for the training and testing of the Intrusion Detection System. Results demonstrated that Random Forest attains the classification accuracy of 99.60% which is the highest in comparison to other machine learning models and the least accuracy of 87.53% has been achieved by Naïve Bayes. Our results also demonstrate that 37.87 seconds is the highest training time required by the Multiple Layer Perceptron whereas the least training time of 0.01 seconds is required by the Naïve Bayes Machine Learning algorithm.

Keywords: Intrusion Detection System, Machine Learning, NSL-KDD dataset, Decision Tree, Naïve Bayes, Random Forest

1.INTRODUCTION

In the rapidly changing advanced world, internet consumption is increasing every day and because of this, the risk of cyber threats also increases quickly. To protect the internet from cyber threats, various cyber security techniques have been used. Cybersecurity is a major concern of this fast-growing technological world [1]. Cyber security is used for the prevention of systems, networks, and data from cyberattacks [2]. Cyber security ensures that all the activities on the computer network must be safe for data and resources. Intrusion on a computer network is one of the major concerns of cyber security. Intrusion Detection System is a major gadget of cybersecurity that is used to prevent illegal intrusion on the network and is defined as a system that continuously monitors every activity and detects any unauthorized activity performed on the computer network. Outdated Intrusion Detection Systems use techniques such as encryption methods, firewalls, and access control mechanisms to detect intrusion on computer networks, but these techniques were not able to provide protection from advanced cyberattacks. Because of the numerous advantages of Deep Learning and Machine Learning techniques, researchers start using various algorithms of Deep Learning and Machine Learning in the Intrusion Detection System for accurate and reliable results. Training time required by Machine Learning algorithms for Intrusion Detection Systems is a matter of concern. A study performed by Platt [3], proposed an express training technique that required less training time for Intrusion Detection Systems techniques using Support Vector Machine algorithms. This model used machine learning techniques for better accuracy to obtain a model for Intrusion Detection. Reduced training time was the main requirement of the study proposed by Platt [3].

2. REVIEW OF RELATED STUDIES

Various researchers used Machine learning for intrusion Detection and obtained better performance than old techniques. In Intrusion Detection System, Feature Selection techniques used also varied according to the dataset. A literature survey on the suitable paper of Intrusion Detection Systems using machine learning was performed and extracts of a few are presented below [4-20].

A method for reducing the training time required by SVM is proposed by Khan et al. [4] so that when dealing with large input datasets, hierarchical clustering analysis can be utilized. For intrusion detection training time is a major concern for the large training dataset, to deal with this concern SVM based IDS system is presented by Khan et al. [4]

Gudadhe et al. [5] presented a model called the hoeffding tree which is a boosted decision tree algorithm used to boost the effectiveness of IDS. Boosting algorithm is used to rise the efficiency of the algorithm by combining a weak base classifier to form a strong classifier. In boosting algorithm, the initially, boosting algorithm started by giving the initial weight of w0 to all data training tuples. Later, on the basis of classification done by the classifier, the load of each tuple is adjusted. After that, another classifier formed the reloaded tuple of training. At last, final categorization of IDS is done on the basis of the average of each classification of all classifiers.

Yasami et al. [6] proposed IDS that can perform unsupervised classification of activities that happen in the network by combining ID3 Decision Tree learning algorithms and K-Means clustering. In this performance matrix evaluations are obtained on the basis of DT and K- means Clustering.

Sangkatsanee et al. [7] combined DT and SVM algorithms so that combined advantages can be used for the IDS. Here, the authors try to improve the results of evaluation parameters by using the pros of Decision Tree and Support Vector Machines algorithms so that model with a better result can be obtained for intrusion detection in the system.

Li et al. [8] proposed an IDS, that works on the RS-FSVM algorithm. Research result shows in the study that RS-FSVM can attain brilliant recognition capacity. In this study, the Authors used an advanced version of SVM known as RS-FSVM to build a model with better accuracy for Intrusion Detection.

Duque et al. [9] Designed an IDS by using K-Means clustering for low false positives, high-efficiency rate, and false negatives model. By using some clusters, K-Means clustering is implemented on the used dataset. In the Scenario when datatypes used in the study matched with a number of cluster matches, idyllic results were obtained in this study.

Aggarwal et al. [10] proposed an IDS in which four categories Host, Basic, Traffic, and Content have 41 attributes of the KDD dataset used to create a dataset of fifteen variants. This newly generated dataset used testing and training of algorithm made by random tree by using the Weka tool. By using the result obtained from the Weka tool, the authors analyzed how every class attribute helps in improving the Detection Rate by minimizing the False Alarm Rate .

Belavagi et al. [11] give a study on the supervised learning algorithm used for IDS and obtained that RF Classifier works best. Here, in this study, various ML algorithms are used and out of that RF performs best for the IDS.

Aburomman et al. [12] proposed a study that used NSL-KDD and KDD99 dataset to review and evaluates the different hybrid techniques and also ensemble techniques on homogeneous and heterogeneous ensemble methods. In this two different NSL-KDD and KDD99 cup dataset was used for the IDS for the required outcome.

Aldwairi et al. [13] designed a system using NSL-KDD that estimates the intrusion scope threshold degree using the best features of network transactions that were made available for training.

Gao et al. [14] construct a Multitree algorithm for an Intrusion Detection System. To increase the detection effect, this study selects a few base classifier algorithms such as Decision Tree, Random Forest, K-Nearest Neighbours, and DNN. By using these models an ensemble adaptive voting model was built to gain the advantages of various algorithms. The accuracy obtained through this method was 85.2%, recall obtained was 85.2%, F1 was 84.9%, and the precision obtained was 86.5% which was better than the individually used base classifier.

Anton et al. [15] proposed a study in which a RF and SVM has been used. 90-95% of the attacks were detected by these algorithms in the dataset. Best predicted result was given by Random Forest in comparison with the Support Vector Machine.

Abrar et al. [16] presented an IDS model on four features extracted from the NSL-KDD dataset using ML algorithms i.e. K-Nearest Neighbours, Support Vector Machine, Naïve Bayes , Decision Tree, Logistic Regression, Random Forest, Multiple Layer Perceptron, and ETC. By using Various subsets of features above 99% performance matrix achieved by Random Forest, ETC, and Decision Tree for attack class classifications.

Sen et al. [17] proposed a study on the specially generated dataset by using J48, Bayes Net, Random Forest, Support Vector Machine, AdaBoost, Multiple Layer Perceptron, Naïve Bayes, and Decision Stump for the detection of DDoS. For this study, a specially generated dataset was used by the researchers.

Husain et. al [18] proposed a study on the CICIDS2017 dataset by using machine learning techniques Multiple Layer Perceptron, Naïve Bayes, Random Forest, Decision Tree, Long Short Term Memory, K-Nearest Neighbours, and J48 for the finding of FTP & SSH Brute-force and various Web cyber-attacks like SQL Injection, Brute-force and cross-site scripting.

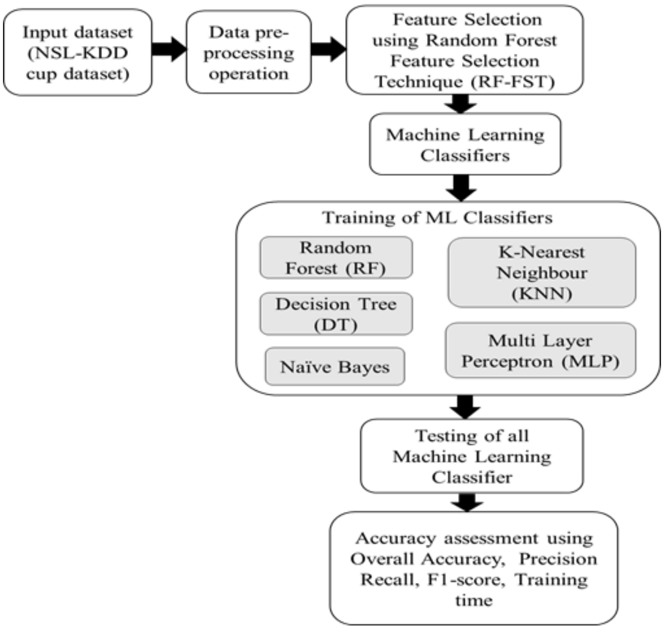
Injadat et al. [19] propose a method on UNSW-NB 2015 and CICIDS 2017 to reduce the computational difficulty of IDS by using the NIDS framework based on ML algorithms. In this study, there was a reduction of 50% & 74% for the feature set and training sample size respectively.

The main agenda of this paper is to identify the best-performing ML algorithms for IDS and also the training time required by the models for IDS on the NSL-KDD dataset.

3. Research Methodology

Here, the Figure shown below represents the workflow diagram of our study. This diagram of the Intrusion Detection System contains steps as Dataset Loading, Pre-Processing, feature Extraction, Train and testing of machine learning algorithms, and calculations of results.

Figure 1. Workflow Diagram


Input dataset (NSL-KDD cup dataset)
Data pre-processing operation
Feature Selection using Random Forest Feature Selection Technique (RF-FST)
Machine Learning Classifiers
Training of ML Classifiers
Random Forest (RF)
K-Nearest Neighbour (KNN)
Decision Tree (DT)
Multi Layer Perceptron (MLP)
Naïve Bayes
Testing of all Machine Learning Classifier
Accuracy assessment using Overall Accuracy, Precision Recall, F1-score, Training time

Firstly, the NDL-KDD dataset is loaded, and then in the pre-processing of the dataset scaling of numerical attributes and encoding of categorical attributes are performed. Whenever raw data is used for research, there is a chance of missing values, null values, and misspelled values. To deal with these abnormalities, pre-processing is a crucial process. In scaling numerical attributes, the characteristics of the dataset are standardized by replacing the mean and scale value of the dataset with unit variance. In label Encoding, the categorical value of the dataset is converted into a number value in the range of 0 to entire class labels minus one. After that, the Random Forest feature extraction technique has been used to choose the best features from the dataset. By using feature selection techniques accuracy of the model increases. The Random Forest-Feature Selection Technique has been used in this study because of its capability to select the best feature in less time as compared to other feature selection techniques.

After that based on training data, Multiple Layer Perceptron, Decision Tree, K-Nearest Neighbours, Random Forest, and Naïve Bayes algorithms are trained. The prediction of unknown data has been done using the termed machine learning models and finally, the proposed algorithms have been compared on the basis of training time, testing time, F1-score, accuracy, precision, and recall. In this study, all implementations have been carried out in Python Programming.

3.1. DATASET DETAILS

A huge dataset is a crucial requirement to train machine learning algorithms. In this paper, the NSL-KDD99 cup dataset has been used for training of Machine Learning model. Since 1999, the KDD-99 dataset is mostly used dataset in cyber security for Intrusion Detection research purposes. It was established by MIT Lincoln Labs. In the KDD-99 cup, training data is collected from five million connections in seven weeks and testing data is collected from 2 million connection records in two weeks. Training data is used by the Machine Learning models for learning patterns and test data on a completely trained solution to analyze its performance. Each connection in the KDD-99 cup is either considered normal or one of four categories of attacks:, Probing, Denial of services, Remote-to-Local (R2L), and User-to-Root. Several glitches of the KDD-99 cup dataset, like unusual data deletion and many more, were removed in NSL-KDD. There are 42 numbers of attributes used in NSL- KDD, in which one is available as a target attribute, and the remaining attributes are taken as input. The used dataset is considered an better version of the KDD-99 cup by removing its cons and hence increasing the training capability of the ML algorithm used for the predictions.

3.2. Performance Metrics

The performance is analysed on the basis of actual and predicted results. The metrics as described below have been considered to analyze the effectiveness of the models.

3.2.1 PRECISION

Minority class accuracy can be calculated using Precision. In Precision only positive prediction is calculated using the following relation:

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positive}}{(\boldsymbol{True\ Positive} + \boldsymbol{False\ Positive})}$$

3.2.2 RECALL

Recall is a measure that determines how many right predictions are obtained from a total number of positive predictions. The expression used to calculate recall is as follows:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positive}}{(\boldsymbol{True\ Positive} + \boldsymbol{False\ Negative})}$$

3.2.3 ACCURACY

It is defined as the percentage of true predictions out of all predictions as presented below:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Negative} + \boldsymbol{True\ Positive}}{\begin{array}{c}(\boldsymbol{True\ Negative} + \boldsymbol{False\ Negative} \\ + \boldsymbol{True\ Positive} + \boldsymbol{False\ Positive}\ )\end{array}}$$

3.2.4 F1-SCORE

The average weight of the Recall and Precision is called as f1-score and that's why both False Positive and False Negative take into consideration for calculating the f1-score using the relation

$$\boldsymbol{f1} = \frac{2(\boldsymbol{Precision} * \boldsymbol{Recall})}{(\boldsymbol{Precision} + \boldsymbol{Recall})}$$

3.2.5 TRAINING & TESTING TIME

Training time described as the time required by the model to learn pattern from the available training dataset for making predictions whereas testing time described as the time required by model to implement learned pattern for the classification of unknown pattern.

4. RESULT AND DISCUSSION

In this unit, results of our proposed work will be discussed. As described earlier,in this work, NSL-KDD dataset has been used for reducing overfitting in the IDS, Random Forest Feature Selection Technique has been used. The top feature selected by Random Forest FST is shown in figure 2.

As shown in above figure, features shown from "rerror_rate" is not important and may cause overfitting. Hence these features of NSL-KDD dataset have not been used for the training of ML classifier model. Also, last 10 least important features are also dropped by us for the better accuracy After feature selection of NSL-KDD dataset, machine learning model has been trained on training dataset. Every machine learning algorithm takes their own training time.

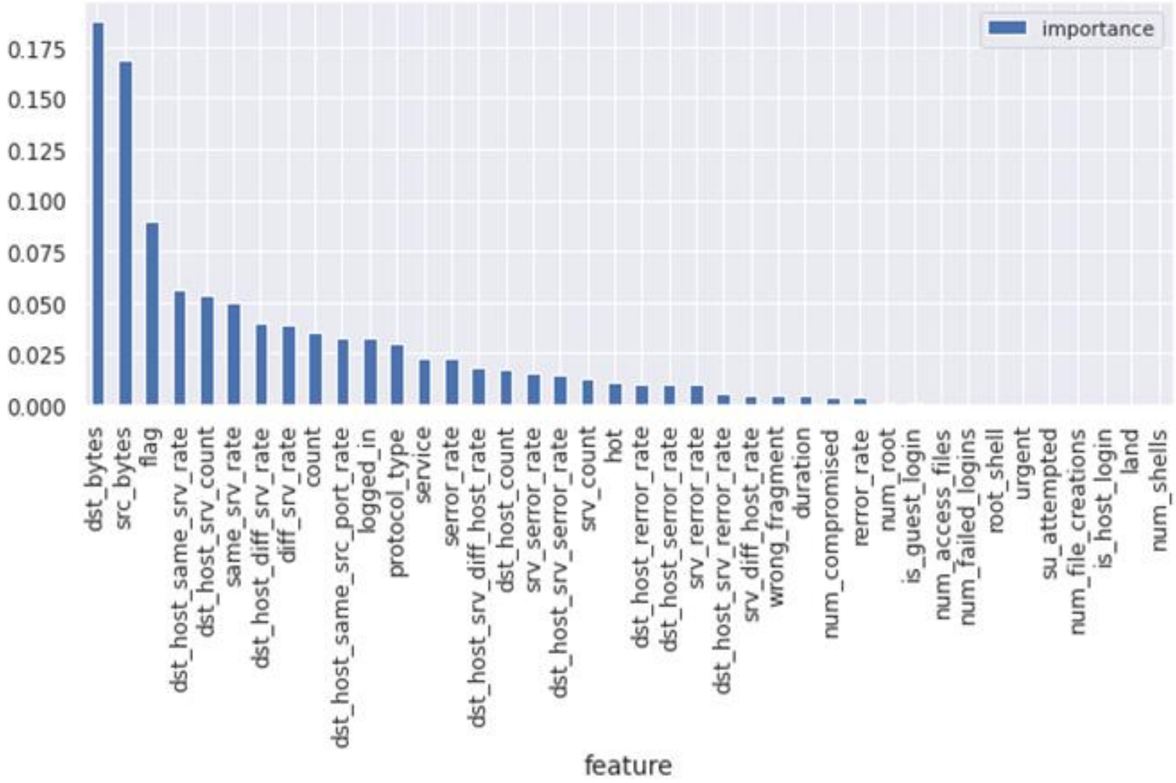


Figure 2. Random-Forest Feature selection Technique

Table 1: Training Time for Machine Learning Algorithms

Machine Learning Algorithm	Training Time (in seconds)
K-Nearest Neighbors	0.03
Multiple Layer Perceptron	37.87
Random Forest	1.35
Naïve Bayes	0.01
Decision Tree	0.06

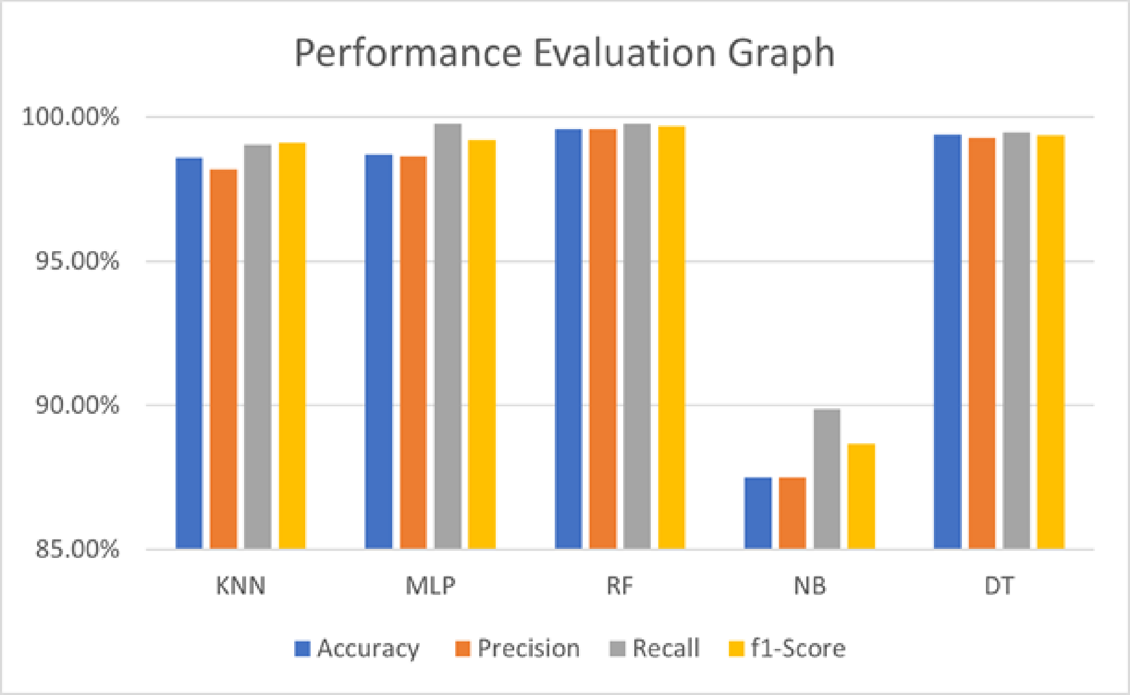
Interpretation of Table 1.

In our study, maximum training time of 37.87 seconds is required by the Multiple Layer Perceptron and least training time of 0.01 seconds is obtained for the Naïve Bayes model. Hence, Multiple Layer Perceptron comes as most time complex model in this study.

Highest Training time of 37.87 seconds is required by Multiple Layer Perceptron model followed by Random Forest(1.35sec), Decision Tree (0.06 sec), K -Nearest Neighbour (0.03 sec) and Naïve Bayes (0.01 sec). From the table we obtained that training time required for Naïve bayes is very small and hence Naïve bayes can be used with large dataset where training time is a limitation.

A comparison graph of performance metrics used for IDS using ML algorithms are given in Figure 3.

Figure 3. Performance Evaluation Graph



Study proposed by [24] used Random Forest, K-Nearest Neighbour, Support Vector Machine, Logistic Regression, and Naïve Bayes for Intrusion Detection System and based on the results, Random Forest observed as best performing algorithm among all of them. However, study proposed by [24] doesn't consider training time parameter to build the model. In our work, results are also same that is Random Forest obtained highest accuracy, but we also considered training time to build the model as important parameter. In our study five algorithms used are K-Nearest Neighbours, Multiple Layer Perceptron, Naïve Bayes, Decision Tree, and Random Forest classifiers based on the results obtained we have observed that although Random Forest is selected as best algorithm in terms of accuracy i.e. 99.60% but training time required by Random Forest is 1.32 seconds which is more than training time of Decision Tree i.e. 0.06 seconds having accuracy of 99.40%.Multiple Layer Perceptron required maximum training time of 37.87 seconds and observed as worst performing algorithm in time critical applications.

6.CONCLUSION

The outcomes of this study revealed that Random Forest is the best-performing algorithm for IDS. The accuracy achieved by Random Forest is 99.6%, which is +12.07% than worst performing classifier i.e., Naïve Bayes. Interestingly, Random Forest and Decision Tree classifiers achieved nearly similar accuracy. Whereas the training time required by Decision Tree is less than Random Forest, hence, in time-critical applications Decision Tree is recommended. Furthermore, in context of time complexity, it has been observed that Multiple Layer Perceptron is the worst performer, whereas Naïve Bayes is best-performing algorithm by consuming minimum time for intrusion detection.

REFERENCES

[1] s. Mukkamala, g. Janoski and a. Sung, "intrusion detection using neural networks and support vector machines," proceedings of the 2002 international joint conference on neural networks. Ijcnn'02 (cat. No.02ch37290), 2002, pp. 1702-1707 vol.2, doi: 10.1109/ijcnn.2002.1007774.

[2] lirim ashiku, cihan dagli,"network intrusion detection system using deep learning, procedia computer science, volume 185,2021,pages 239-247,issn 1877-0509,https://doi.org/10.1016/j.procs.2021.05.025.
[3] j.platt, "fast training of support vector machines using sequential minimal optimization", advances in kernel methods: support vector learning, pp. 185-208, 1998.

[4] l.khan, m.awad, and b.thuraisingham. "a new intrusion detection system using support vector machines and hierarchical clustering" the vldb journal—the international journal on very large data bases 16, no. 4 (2007): 507-521.

[5] m.gudadhe, p.prasad, and k.wankhade. "a new data mining based network intrusion detection model" international conference on computer and communication technology (iccct), pp. 731-735. Ieee, 2010.

[6] y.yasami and s.p.mozaffari. "a novel unsupervised classification approach fornetwork anomaly detection by k-means clustering and id3 decision tree learning methods" the journal of supercomputing 53, no. 1(2011).231-245.

[7] p.sangkatsanee, n. Wattanapongsakorn and c. Charnsripinyo, "practical real -time intrusion detection using machine learning approaches, computer communications" , vol. 34, no. 18, pp. 2227– 2235, (2011).

[8] l.li, and k.n zhao, "a new intrusion detection system based on rough set theory and fuzzy support vector machine" 3rd international workshop on intelligent systems and applications (isa), 2011, pp. 1-5. Ieee, 2011.

[9] s.duque and n.b omar. "using data mining algorithms for developing a model for intrusion detection system (ids)" procedia computer science 61 (2015), pp. 46–51.

[10] p.aggarwal and s.sharma, "analysis of kdd dataset attributes - class wise for intrusion detection" procedia computer science (2015) 57. 842-851. 10.1016/j.procs.2015.07.490.

[11] m.c. belavagi and b.muniyal. "performance evaluation of supervised machine learning algorithms for intrusion detection" twelfth international conference on communication networks, iccn 2016, august 19– 21, 2016.

[12] a. A. Aburomman and m. B. I. Reaz, "a survey of intrusion detection systems based on ensemble and hybrid classifiers" comput. Secur., vol. 65, pp. 135– 152, 2017.

[13] aldwairi, and m. B. Yassein,"anomaly - based intrusion detection system throughfeature selection analysis and building hybrid efficient model," j. Comput. Sci., vol. 25, pp. 152– 160, 2018.

[14] x. Gao, c. Shan, c. Hu, z. Niu and z. Liu, "an adaptive ensemble machine learning model for intrusion detection," in ieee access, vol. 7, pp. 82512-82521, 2019, doi: 10.1109/access.2019.2923640.

[15] s. D. D. Anton and s. Sinha. "anomaly-based intrusion detection in industrial data with svm and random forests". In: 2019 international conference on software, telecommunications and computer networks (softcom). 2019, pp. 1–6. Doi: 10.23919/softcom.2019.8903672.

[16] i. Abrar, z. Ayub, f. Masoodi and a. M. Bamhdi, "a machine learning approach for intrusion detection system on nsl-kdd dataset," 2020 international conference on smart electronics and communication (icosec), 2020, pp. 919-924, doi: 10.1109/icosec49089.2020.9215232.

[17] s. Sen, k. D. Gupta, and m. M. Ahsan, "leveraging machine learning approach setup software-defined network (sdn) controller rules during ddos 223 attack." In proceedings of international joint conference on computational intelligence. Springer, 2020, pp. 49–60. Https://doi.org/10.1007/978-981-13-7564-45.

[18] m. D. Hossain, h. Ochiai, f. Doudou, and y. Kadobayashi, “ssh and ftp brute-force attacks detection in computer networks: lstm and machine learning approaches.” In 5th international conference on computer and communication systems (icccs). Ieee, 2020, pp. 491–497. Https://doi.org/10.1109/icccs49078.2020.9118459.

[19] m. Injadat, a. Moubayed, a. B. Nassif, and a.shami,"multi-stage optimized machine learning framework for network intrusion detection," in ieee transactions on network and service management, vol. 18, no. 2, pp. 1803-1816, june 2021, doi: 10.1109/tnsm.2020.3014929.

[20] guo, gongde & wang, hui & bell, david & bi, yaxin. (2004). Knn model-based approach in classification.

[21] mukherjee, saurabh & sharma, neelam. (2012). Intrusion detection using naive bayes classifier with feature reduction. Procedia technology. 4. 119–128. 10.1016/j.protcy.2012.05.017.

[22] patel, harsh & prajapati, purvi. (2018). Study and analysis of decision tree based classification algorithms. International journal of computer sciences and engineering. 6. 74-78. 10.26438/ijcse/v6i10.7478.

[23] nazzal, jamal & el-emary, ibrahim & najim, salam. (2008). Multilayer perceptron neural network (mlps) for analyzing the properties of jordan oil shale.world applied sciences journal.

[24] b.yogesha, dr. G. Suresh reddy, " intrusion detection system using random forest approach," turkish journal of computer and mathematics education, vol.13 no.02, pp. 725-733,2022.