# Analysis of IDS using Feature Selection Approach on NSL-KDD Dataset

Rahila Rahim, Aamir S Ahanger, Sajad M Khan, Faheem Masoodi

Department of Computer Science, University of Kashmir, India

Corresponding author: Faheem Masoodi, Email: masoodifahim@uok.edu.in

Due to the increased use of the internet, cyber-attacks are becoming more prominent causing major difficulty in achieving and preventing security risks and threats in the network.There have been a variety of attacks (both passive and aggressive) used to compromise network security and privacy. For detecting such attacks quickly and accurately, a strong Intrusion Detection System is required which is a valuable means for detecting intrusions in a network or system by extensively inspecting each packet in the network in real-time, preventing any harm to the user or system resources. In this paper, we proposed a statistical method to train the model with the training data to understand complicated patterns in the dataset and to make intelligent decisions or predictions whenever it comes across new or previously unseen data instances. For the classification of data, we used five machine learning classifiers such as Support Vector Machine, Decision Tree, Random Forest, AdaBoost, and Logistic Regression. To properly grasp complicated patterns in data, machine learning models require a large amount of data, which is why NSL-KDD was utilized to develop and validate supervised machine learning models. Initially, the dataset is pre-processed to remove any unnecessary or undesired dataset features. Feature selection (extra-tree classifier) were used which combines the qualities of both filter and wrapper methods to provide features based on their importance as a result, the dataset dimensionality is reduced, lowering the processing complexity. Finally,the overall classification accuracy of the various machine learning classifiers was evaluated to find the best optimal algorithm for detecting intrusions.

**Keywords**: Intrusion detection, NIDS, Feature selection, Machine learning.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 475–481. Computing & Intelligent Systems, SCRS, India.
https://doi.org/10.52458/978-93-91842-08-6-45

## 1 Introduction

Network security is defined as a set of standard laws and practices adopted by network officials to protect the integrity, confidentiality, and accessibility of networks and their resources from unauthorized access, misuse, modification, or denial of service through the use of various security mechanisms such as firewalls, anti-virusor, anti-malware software, data encryption techniques, and IDS is widely regarded as the most effective method for jeopardizing systemsecurity [1]. An Intrusion detection system (IDS) is a type of network security technology used to identifying attacks in the target systems that are accessible. It's a piece of software that examines system communications for nefarious activities and rule violations. Any heinous crime is frequently reported to network administrators using a security information and event management (SIEM) system, which raises alarm when such activity is identified. Network-based intrusion detection system (NIDS) and host-based intrusion detection system (HIDS) are the two types of IDS based on their deployment (HIDS) [2]. NIDS often include a hardware component (sensor) as well as a network interface card (NIC) that is installed on the network to detect malicious traffic. It examines incoming network traffic from numerous sources in promiscuous mode and compares it to a database of known threats in the system, sending an alarm to the NIDS management server if such activity is detected, whereas HIDS is installed on a host machine [3]. It monitors the system's incoming and outgoing traffic for intrusions and policy violations. It compares the system file to the existing database, then logs the action and alerts the network management if suspicious or malicious behavior is identified. Misuse or SIDS (signature-basedIDS), AIDS (Anomaly-based IDS), and HIDS (hybrid-based IDS) are three regularly utilized methodologies in intrusion detection [1]. Signature-based IDS examines a unique set of patterns noted as sequences for identifying potential threats. This approach scans and compares incoming traffic against the available signatures by keeping a database of attack signatures/Rules frompreviously known assaults in the system. When signatures are matched, an alert is sounded. Because there is no accessible sequence, the SIDS can identify known attacks smoothly but cannot detect new attacks. Anomaly-based IDS (AIDS) is used to identify unidentified assaults. It's a statistical model that employs machine learning techniques to create a reliable activity model for detecting unknown malware threats. Hybrid-based IDS is created by integrating twointrusion detection system approaches (AIDS and HIDS) [4]. It identifies as many threats as possible and tries to avoid all of the strategies used by intruders to obtain access to the system (fragmentation, spoofing, and so on). Machine learning is a crucial component of any growing field of data science. It employs statistical approaches and algorithms to train the model based ontraining data. When it encounters unseen data of its own, it pulls unique and hidden patterns fromthe dataset to create predictions or classifications and predict future trends from the data [5]. Two types of Machine learning algorithms are: unsupervised and supervised learning. During training phase of Supervised Learning, the model is trained with the dependent variable (data along with outcome), whereas in Unsupervised Learning, the model is learned on unlabeled data (data with input butno output) [6]. The model learns through observations and groups' unsorted data into groups based on similarities, patterns, and differences, or the data's intrinsic qualities. Many machine learningclassifiers were employed in intrusion detection systems and feature selection to detect intrusions effectively and with high accuracy rates including Decision trees, RF, SVM, logistic regression, XGBoost. [7][8].

## 2 Literature Survey

Iram et al. [9] employed a variety of machine learning algorithms to determine if the data wasnormal or malicious, including LR, SVM, K-nearest neighbor, Nave Bayes, Extra-tree classifier, MLP, DT, RF. The NSL-KDD dataset was used to create multiple feature subsets, from which the empiricalresult of RF, ETC, and DT comes to be greater than 99 percent. As a result, the suggested modelhas a high accuracy rate while also reduces computational complexity by removing irrelevant data.

Ahanger et al [10] used SVM, RF MLP, and DT totest and trains the model. To detect the intrusion effectively with greater accuracy and performance, feature choices were used to remove the dataset's irrelevant and undesirable features.

Ahmad et al [11] used supervised ML algorithms such as RF, SVM, and extreme learning machine (ELM) for classification problems, which results in better performance than other classifiers such as multilayer perceptron, naive Bayes, self-organizing map, and others, which do not work well on large datasets and reduce detection rate. The evaluation findings suggest that ELM outperforms SVM in terms of detection and SVM in turnout performs RF in terms of outcomes.

AdaBoost, a machine learning classifier in which both categorical and continuous features arelined by decision rules, is used by Hu et al [12] to detect intrusions in the network. Later, both the categorical and continuous weak classifiers (SVM, KNN) are combined into astronger classifier, which classifies the data more accurately than the weak classifier. Adjustable initial weights and a basic strategy to avoid overfitting are used to improve the algorithm'sperformance.

On the KDD99 dataset, Tahir et al. [13] constructed an AIDS based on several ML algorithmssuch as SVM, nave-bayes, j.48, decision table, and compared these algorithms to each other in order to detect traffic behavior effectively with the fewest false positives. Decision tree and j.48 yield the best results, with a detection rate of more than 99 percent.

Anisa et al [14] used two different ML algorithms (naive bayes and SVM) to create IDS that was more accurate and had a higher detection rate. The results demonstrate that SVM outperforms the Nave Bayes classifier by a factor of 97 percent.

Sumaiya et al [15] use Wire shark to collect network traffic packets online and store them in adatabase. Later, use various machine learning methods (SVM, RF, nave bayes, KNN) to detectnetwork breaches. Finally, Random Forest outperforms the other methods, with a maximumaccuracyof 99.81 percent.

# 3 Material and Methods

## A. Dataset Preprocessing

NSL-KDD was used to examine the model's performance in this methodology. NSL-KDD is a more concise & efficient intrusion detection dataset developed by the University of New Brunswick in which the records were carefully selected. The dataset consists of a total 43 features, out of which 41 features are input traffic and 2 attributes consist of labels i.e., label and score (labeled fines multiclass attacks) and Score (defines the sharpness of the normal & malicious traffic) [12]. Data-preprocessing is an important stage through which the unprocessed data is converted intothe processed and readable format. Preprocessing helps in achieving the best results to create a machine learning model for training purposes [16]. Preprocessing minimizes computational complexity while improving generalization ability. In the NSL-KDD dataset, we have three features of categorical type namely Service, Protocol, and Flag. Before training, testing the models these categorical data were transformed in to integer types using the Label Encoding technique. Some feature values in the data set have a very larger range which has a negative influence on the performance of the models. To eliminate the adverse effect, we employed min-max normalization.

## B. Feature Selection

The 2nd phase is the feature selection phase that includes selecting those features which are most important and dropping those which are less important. The purpose of using this phase is to provide optimal results and to increase the classification accuracy by shrinking the dimensionality of a dataset so that the Machine Learning algorithms perform classification more efficiently. In the project, we employed the embedded method (Extra tree classifier) for feature selection which combines the qualities of both filter and wrapper methods and provide features based on their importance, the important attributes identified by extra tree classifier are presented in the table.1

**Table 1.** Important Attributes Identified using Feature Selection

| S. No. | Name of the Selected Features |
|---|---|
| 1. | Service |
| 2. | Flag |
| 3. | logged_in |
| 4. | Count |
| 5. | serror_rate |
| 6. | srv_serror_rate |
| 7. | same_srv_rate |
| 8. | dst_host_serror_rate |
| 9. | dst_host_srv_serror_rate |
| 10. | dst_host_srv_count |
| 11. | dst_host_rerror_rate |
| 12. | dst_host_same_srv_rate |
| 13. | dst_host_diff_srv_rate |
| 14. | dst_host_same_src_port_rate |
| 15. | dst_host_serror_rate |

## C. Classification

Various machine learning classifiers were implemented in this paper to make a better model which classifies the data as normal or malicious with maximum precision and a low false alarm rate. Machine learning algorithms like decision trees, random forests, AdaBoost classifier, logistic regression, and support vector machines were used. The supervised machine learning method decision tree is used to solve both regression and classification issues. The aim of employing a decision treecreates a teaching model that simply applies thedecision rules learned in the training set to determine the classification result of the feature [17]. Decision tree (DT) employs a recursive method to categorize each class and to predict the classlabel initiated from root node and correlates the values from the root attribute with the values they contain in the attribute records. Random forests (RF) use the concept of ensemble learning which merges multiple decision trees to perform the complex problems and to increase the model's efficiency for better prediction [18]. Random forest finds the predictions of each decision tree and chooses the result based on the majority votes of predictions, to classify the data accurately. Logistic Regression (LR) is an instance of supervised learning method that is commonly adopted for classification purposes. It helps to predict the probabilities of the possible target variable (dependent variable) with one or more features used to predict the target variable (in dependent features) using a logistic function [19]. Support vector machine (SVM) is a classification technique used for both linear and non-linear data but mostly suited for non-linear data separation problems [20]. The SVM algorithm's main goal is to divide datasets into classes so that an optimal marginal hyperplane (MMH) can be found between the extreme points [21]. Ada-boost, or Adaptive Boostin g is an ensemble boosting that works on the process of voting. This algorithm is used to check the model which doesn’t result in accurate prediction. It creates iterative ensembles that combine many low-performing classifiers to improve classifier accuracy. Its primary purpose isto construct classifier weights in which the data are trained at each stage to make accurate predictions for uncommon occurrences [22]. Figure 1 depicts the flow chart for the planned work.

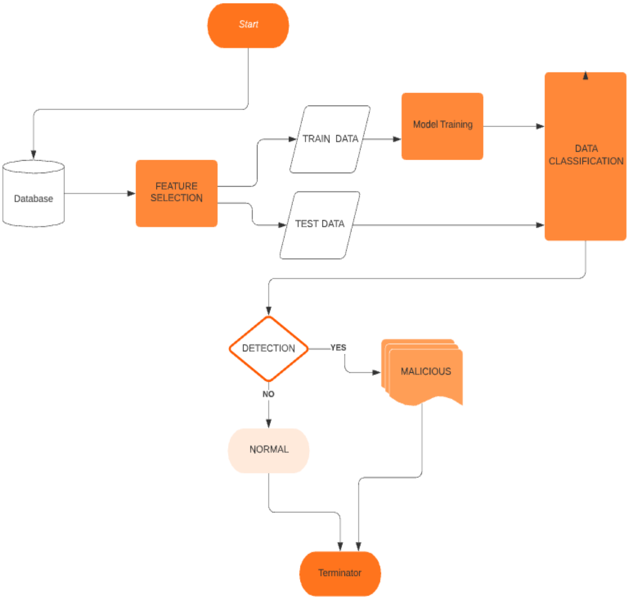


**Fig 1.** Flow chart of the classification process

## 4 Results

The experiment was conducted on Google Colab using various machine learning models. To train and test the models, the feature subset was selected using Extra Tree Classifier from the original dataset to reduce the number of dimensionalities of the data and to increase the detectionrate. The various classification models including DT, RF, LR, SVM, AdaBoost were trained and tested using the NSL-KDD dataset. The accuracy achieved by the proposed classifiers is displayed below in figure2. and table2.The result shows that the overall accuracy of decision tree, random forest and logistic regression comes to be greater than 99 percent.

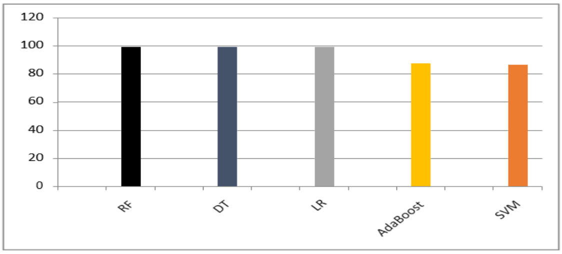
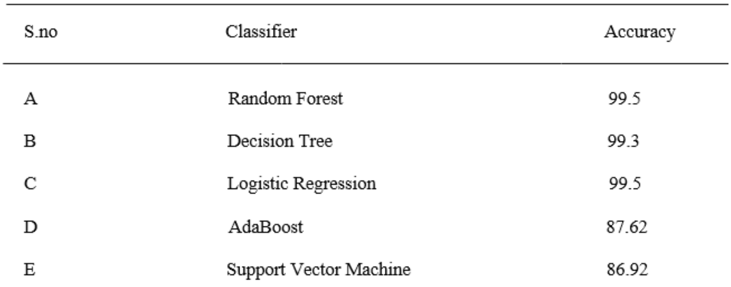


**Fig. 2.** Accuracy of the different ML Approaches

**Table 2.** Accuracy of the ML Approaches



| S.no | Classifier | Accuracy |
|---|---|---|
| A | Random Forest | 99.5 |
| B | Decision Tree | 99.3 |
| C | Logistic Regression | 99.5 |
| D | AdaBoost | 87.62 |
| E | Support Vector Machine | 86.92 |

## 5 Conclusion and Future Work

In the project, experimental analysis was performed to test and assess the efficiency and effectiveness of various machine learning classifiers which include DT, RF, LR, SVM, and AdaBoost on NSL-KDD dataset for detecting intrusions. To enhance effectiveness and to decrease training time, the dataset was first preprocessed along with the selection of important features using feature selection method. As aresult of the classifiers we utilized, we were able to generate superior outcomes. The overall classification rate of the DT, RF, and LR models is more than 99 percent. In the future, intrusion detection can be performed on a real-time dataset utilizing a combination of algorithms to detect unknown threats with increased model accuracy and lower false detection rates.

# References

[1] Khraisat, A. et al. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cyber Security,* 2:20.

[2] Liao, H. J. et al. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications,* 36(1):16–24.

[3] Bamhdi, A. M., Abrar, I. and Masoodi, F. (2021). An ensemble based approach for effective intrusion detection using majority voting. *Telecommunication Computing Electronics and Control*, 19(2):664–671.

[4] Bokhari, M. U. and Masoodi, F. (2012). BOKHARI: A new software oriented stream cipher: A proposal. In *Proceeding of World Congress on Information & Communication Technololgy,* 128–131.

[5] Alazzam, H., Sharieh, A. and Sabri, K. E. (2020). A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer. *Expert Systems with Applications*, 148:113249.

[6] Masoodi, F., Alam, S. and Bokhari, M. U. (2011). SOBER Family of Stream Ciphers: A Review. *International Journal of Computer Applications*, 23(1):1–5.

[7] Fernandes, G. et al. (2019). A comprehensive survey on network anomaly detection. *Telecommunication Systems*, 70(3):447–489.

[8] Almasoudy, F. H., Al-Yaseen, W. L. and Idrees, A. K. (2020). Differential Evolution Wrapper Feature Selection for Intrusion Detection System. *Procedia Computer Science*, 167:1230–1239.

[9] Abrar, I. et al. (2020). A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. In *Proceeding of the International Conference of Smart Electronics Communication*, 919–924.

[10] Ahanger, A. S., Khan, S. M. and Masoodi, F. (2021). An Effective Intrusion Detection System using Supervised Machine Learning Techniques. In *Proceeding of 5th International Conference on Computing Methodologies and Communication*, 1639–1644.

[11] Ahmad, I. et al. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme

Learning Machine for Intrusion Detection. *IEEE Access*, 6:33789-33795.

[12] Hu, W. et al. (2008). AdaBoost-Based Algorithm for Network. *IEEE Transactions on Systems, Man, and Cybernetics,* 38(2):577–583.

[13] Mehmood, T. and Rais, H. B. M. (2016). Machine learning algorithms in context of intrusion detection. In the *3rd International Conference on Computer and Information Sciences,* 369–373.

[14] A. H. A. and Sundarakantham, K. (2019). Machine Learning Based Intrusion Detection System. *In* the *2019 3rd International Conference on Trends in Electronics and Informatics,* 916–920.

[15] Thaseen, I. S., Poorva, B. and Ushasree, P. S. (2020). Network Intrusion Detection using Machine Learning Techniques. In the *International Conference on Emerging Trends in Information Technology and Engineering,* 1–7.

[16] Masoodi, S. T. et al. (2019). Security and privacy threats, attacks and countermeasures in Internet of Things. *International Journal of Network Security & Its Applications,* 11(2):67–77.

[17] Ahmed, F. et al. (2021). Security Concerns and Privacy Preservation in Blockchain based IoT Systems: Opportunities and Challenges. In the *International Conference on IoT Based Control Networks and Intelligent Systems,* 29–36.

[18] Pandow, B. A., Bamhdi, A. M. and Masoodi, F. (2020). Internet of Things: Financial Perspective and Associated Security Concerns. *International Journal of Computer Theory and Engineering*, 12(5):123–127.

[19] Masoodi, F. et al. (2021). Machine Learning for Classification analysis of Intrusion Detection on NSL-KDD Dataset. *Turkish Journal of Computer and Mathematics Education,* 12(10):2286–2293.

[20] Masoodi, F. S. and Bokhari, M. U. (2019). Symmetric Algorithms I. *Emerging Security Algorithms and Techniques,* 79–95.

[21] Teli, T. A. and Masoodi, F. (2021). Blockchain in Healthcare: Challenges and Opportunities. *SSRN Electronic Journal*, 1–6.

[22] Abrar, I. et al. (2020). A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. *In the International Conference on Smart Electronics and Communication,* 919–924.