Machine Learning Model for Intrusion Detection

Devishree Naidu, Sandesh Sachdev and Vaidik Murarka

November 18, 2022

Machine Learning Model for Intrusion Detection

1st Devishree Naidu
Department of Computer Science and Engineering
Shri Ramdeobaba College of Engineering and Management
Nagpur 440013,India
naidud@rknec.edu

2nd Sandesh Sachdev
Department of Computer Science and Engineering
Shri Ramdeobaba College of Engineering and Management
Nagpur 440013,India
sachdevst@rknec.edu

3rd Vaidik Murarka
Department of Computer Science and Engineering
Shri Ramdeobaba College of Engineering and Management
Nagpur 440013,India
murarkavm@rknec.edu

***Abstract*—In the field of network security, there is a never-ending search for cyber-attacks that might disrupt a network. Furthermore, with the unanticipated emergence and expanded use of the Internet, hostile network activities are rapidly increasing. It is critical to build a comprehensive intrusion detection system (IDS) to combat unwanted access to network resources in order to detect anomalies in the network and secure information. Intrusion Detection System (IDS) has been an efficient technique to attain improved security in identifying harmful activity.**

Because it is unable to detect all sorts of attacks correctly, current anomaly detection is frequently linked with high false alarm rates and only modest accuracy and detection rates. Intrusion detection systems search for signatures of known attacks or abnormal activities. Machine learning approaches are taken to approach in this project by using the KDD-99 Cup and NSL-KDD datasets, experiment is conducted to evaluate the performance of several machine learning methods.

Using the NSL-KDD datasets, an experiment is conducted to analyse the effectiveness of several machine learning methods in order to design a methodology for creating a Machine Learning Modal with a higher prediction rate in detecting an attack on the host network. The results reveal which method worked best in terms of accuracy, detection rate, and false alarm rate.

The performance of RF, KNN for all attack classes utilising different feature subsets was above 99 percent. As a result, the suggested model has a high accuracy rate while also reducing computational complexity by eliminating unimportant elements.

***Index Terms*—NSL-KDD,IDS,ML,AirGap Security**

I. Introduction

When it comes to Network Security, No such device is perfectly secure. One that is connected to the Internet is definitely not. According to McAfee, Hackers create 300,000 new pieces of malware daily. Many Businesses and Big Organizations spend millions of dollars trying to protect their sensitive information from the reach of hackers who could destroy the company's business plan, because of which they try to Air gap their network to safeguard from potential threats. But this is not enough as its loopholes have been constantly being discovered and thus increasing. This gave us the motivation to develop a modal based on Intrusion detection which will detect whether the breach was conducted on the system or not, which will help to fix the bugs as earlier as possible.

II. Related Work

NSL KDD dataset is an improved version of its predecessor, KDD'99. This report analyzes and uses the NSL KDD dataset, Different classification algorithms for in detection Abnormal Network traffic patterns. Intrusion detection analysis has become very important in the last decade.[9] Based on DAPRA 98, researchers are focusing on different datasets to improve the accuracy of the system and reduce false positives. Anomaly-based attack detection systems using machine learning techniques can be trained for detection. Current anomaly detection is often associated with high false alarm rates and moderate accuracy and detection rates of , as not all types of attacks can be detected correctly. The experimental results obtained show that the proposed method achieved 91% classification accuracy using only 3 features and 99% classification accuracy using all 36 features and 41 training features. It shows that you did. Indicates. It shows that 99% classification accuracy has been achieved.

A. Already Proposed Explorations on NSL KDD

1) A Survey on Machine Learning based Intrusion Detection System on NSL-KDD Dataset: In this paper,[10] a literature survey has been conducted upon different research papers who have used either used KDD 99' or NSL-KDD security dataset to study upon the intrusion detection system or on the idea to implement a IDS model using machine learning techniques. A number of selection techniques and classifiers including KNN, K Means SVM also work on Deep Belief Network, Genetic algorithm and DCNN (Deep Convolution Neural Network) has been done on this security data set. This research works concludes to work on the hybrid approach for making the development to the intrusion detection technology in machine learning domain.

2) Performance Analysis of NSL-KDD dataset using ANN: The research paper proposed to apply the Artificial Neural Network on the NSL-KDD dataset.[5] The proposed neural network architecture used are tansig transfer function, Levenberg-Marquardt (LM) and BFGS quasi-Newton Backpropagation (BFG) algorithm. The learning in ANN is performed by changing the values of neurons and layers. The proposed techniques in the paper proposed that for binary class classification, it gives higher accuracy of attack detection than that of other reported technique. For five class classification it was found that the system has good capability to find the attack for particular class in NSL-KDD dataset.

3) Feature Selection and Intrusion classification in NSL-KDD Cup 99 Dataset Employing SVMs: The proposed research paper uses SVM's (Support Vector Machines). SVM's are powerful machine learning algorithm that are for pattern recognition, image classifications and biometrics analysis.[7] The proposed method keeps the SVM classifier's classification accuracy but employs a smaller number of input characteristics from training data. The NSL-KDD Cup 99 dataset contains normal and attack network connections and is a multiclass classification problem. The suggested method achieved 91 percent classification accuracy using only three input features and 99 percent classification accuracy using 36 input features, while all 41 input features achieved 99 percent classification accuracy in this study.

B. Methodology

Taking all the literature survey and objectives in mind, the following methodology was decided upon for this project.

III. DATASET

[8]NSL-KDD is an improved version of the KDD cup99 data set that addresses some of the issues with the prior version. Many different sorts of analyses have been performed on the NSL-KDD dataset by many researchers using various methodologies and tools with the common goal of developing an effective intrusion detection system.

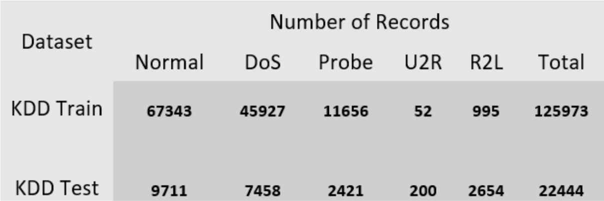
[2] This data set is based on the DARPA 1998 data set created by the Cyber Systems and Technological group of the MIT Lincoln laboratory.

The WEKA programme does a deep study of the NSL-KDD data set using multiple machine learning approaches. The NSL-KDD data set is used to train and test different existing and new attacks using the K-means clustering algorithm. Many different sorts of analyses have been performed on the NSL-KDD dataset by many researchers using various methodologies and tools with the common goal of developing an effective intrusion detection system. The NSL-KDD data set is used to train and test different existing and new attacks using the K-means clustering algorithm.

[4][**?**] NSL-KDD data set is a refined version of its predecessor. It contains essential records of the complete KDD data set. There are a collection of downloadable files at the disposal for the researchers. Although the NSL-KDD data set suffers from some problems, it is a very effective data set that can be used for research purposes [1] [4]

TABLE I
LIST OF NSL-KDD DATASET FILES AND THEIR DESCRIPTION

Sr. No.	Name of the File	Description
1	KDDTrain+.ARFF	The full NSL-KDD train set with binary labels in ARFF format
2	KDDTrain+.TXT	The full NSL-KDD train set including attack-type labels and difficulty level in CSV format
3	KDDTrain+$_2 0 Percent.ARFF$	A 20% KDDTrain+.arff file
4	KDDTrain+$_2 0 Percent.TXT$	A 20% KDDTrain+.txt file
5	KDDTest+.ARFF	The full NSL-KDD test set with binary labels in ARFF format
6	KDDTest+.TXT	The full NSL-KDD test set including attack-type labels and difficulty level in CSV format
7	KDDTest-21.ARFF	A subset of the KDDTest+.arff file which does not include records with difficulty level of 21 out of 21
8	KDDTest-21.TXT	A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21



Dataset	Number of Records					
	Normal	DoS	Probe	U2R	R2L	Total
KDD Train	67343	45927	11656	52	995	125973
KDD Test	9711	7458	2421	200	2654	22444

Fig. 1. The Fig. show the data records in the NSL-KDD Dataset.

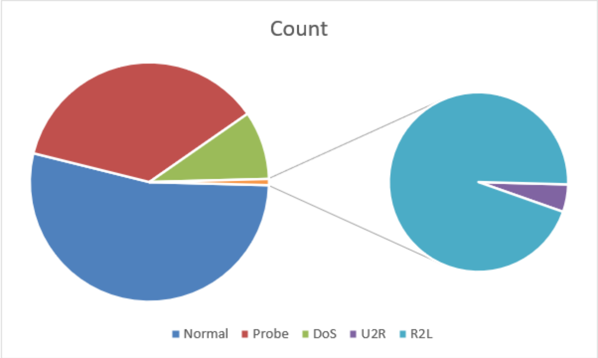


Fig. 2. The Fig. show the data records in the NSL-KDD Dataset on attacks class in the form of Pie Chart.

Four special types of attacks are present in the Dataset: Denial of Service (DoS), Probe, User to Root(U2R), and Remote to Local (R2L). A short description of every attack is described below:

1) DOS: Denial of Service attacks are done with the aid of using flooding target hosts or networks with carefully dependent traffic in a way calculated to deplete the

targets' sources and thereby both suspend or crash their operations. – e.g. syn flooding.

2) Probing: A probe is a software inserted at a key juncture in a network for the motive of tracking or collecting information about network activity.The attacker gathers statistics about the shape of the assisting infrastructure.
3) U2R: A User to Root (U2R) attack the target system first succeeds in gaining a foothold at the remote system withinside the shape of a consumer session, preferably withinside the form of an interactive shell. By combining a whole lot of traditional techniques, the attacker endeavors incrementally to escalate his privileges till he achieves super-user permissions.
4) R2L: A Remote to User attack is conceptually similar to the user-to-root attack however is modest in its ultimate ambition. Such an attack is transacted while an attacker sends packets to the target host machine which might be intended to disclose vulnerabilities that might permit the attacker to take advantage of a nearby user's privileges.

IV. Machine Learning Processes

A. Data Preprocessing

It is a step in the data mining and data evaluation procedure that takes raw information and transforms it right into a format that may be understood and analyzed through computers and machine learning. Raw, real-world information in the shape of text, photos is messy and may incorporate errors and inconsistencies, however it's far frequently incomplete, and doesn't have a regular, uniform design.

B. Data Visualization

It is the step in machine learning to graphically visualize the data by the means of pie-charts, bar graphs, scatter plots etc.This may be beneficial while exploring and gaining knowledge of a dataset and might assist with figuring out patterns, corrupt records, outliers, and much extra. With a bit domain knowledge, records visualizations may be used to specific and exhibit key relationships in plots and charts which might be extra visceral to your self and stakeholders than measures of association or significance.

C. Feature Engineering

It is a machine learning approach that leverages information records to create new variables that aren't in the training set.These artificial functions are then used by that set of policies as a manner to enhance its performance.It consists of creation, transformation, extraction, and choice of functions, additionally known as variables.These Process are:

1) Feature Creation : Creating functions entails figuring out the variables to be able to be maximum beneficial withinside the predictive version. This is a subjective procedure that calls for human intervention and creativity.
2) Transformation : Transformation entails manipulating the predictor variables to enhance version performance.
3) Feature Extraction : Feature extraction is the automated creation of recent variables with the aid of using extracting them from raw statistics. The cause of this step is to automatically reduce the extent of data right into a extra manageable set for modeling
4) Feature Selection : It is a procedure getting rid of low accuracy functions for the intention of having quicker training time. The procedure selects the attributes which can be fairly affecting the outcome/prediction from the model.

D. Random Forest Classifier

It includes generation of decision trees on different samples and then analyze the outcomes, giving the result on the majority of votes from the decision trees. Instead on just relying on a decision tree itself, random forest creates vast number of tress to improve its performance.It's more accurate than the decision tree algorithm.It provides an effective way of handling missing data.It can produce a reasonable prediction without hyper-parameter tuning.It solves the issue of overfitting in decision trees.In every random forest tree, a subset of features is selected randomly at the node's splitting point.

E. KNN Classifier

It classify data points based on similarities. It is commonly referred to as a lazy algorithm because it does not develop a learning algorithm to predict the outcome. It do not assume how the model will be created from the given data. This is useful when performing pattern recognition tasks that classify objects based on different characteristics.

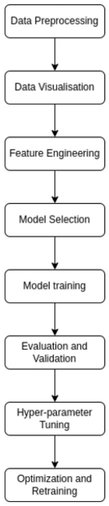


Fig. 3. The Figure show different phases in the machine learning development.

V. DATA PREPROCESSING

Before Importing the dataset, Some Functional libraries have to be imported to make the task to machine learning classifications and regression functions readily available to us. These include Pandas, Numpy, Matplotlib, Seaborn, SK Learn.

The NSL KDD dataset has an advantage over KDD 99' that this has much cleaner records and so no such pre-processing was much required to it.

VI. Exploratory Data Analysis

A. Encoding

Upon exploration, the dataset is found to have 42 attributes, out of which 41 are features that are determining the packet information. And the last feature in the dataset is actual class of the attack. Out of all the features in the records, 3 features were of type object (string) while others being int or float values. The object values have to be converted into numerical ones, so that the classification models could work on it. So, we converted these object values using hot-encoding and label encoding wherever required.

B. Data Visualization

In this, we prepared some graphical presentation to pictorially represent the records from the dataset.Let's take a look at some charts to see how things are distributed.

The thing to notice here is the difference in each protocol type. Our initial impression is that protocol may be useful in being able to identify the type of traffic we are observing. Let's see if flag behaves the same way.

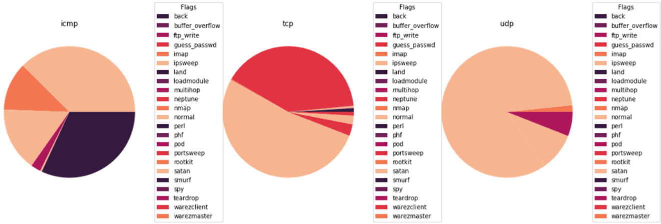


Fig. 4. The Fig. shows the dataset records distinguished on the basis of transport layer protocol by the attack they portrayed.

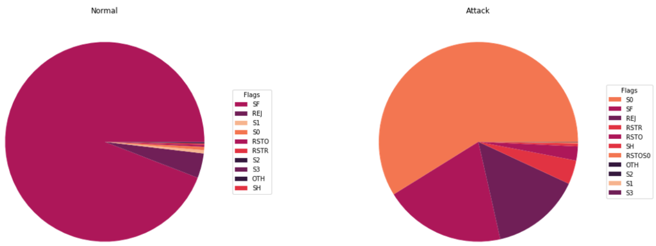


Fig. 5. The Fig visualises Normal and Attack packets on the flags active on the packet.

If we think about this from the eyes of a network adminstrator, the combination of protocol, flag and service seem like they should tell us a lot about the nature of our traffic.

C. Feature Engineering

So let's dive into some feature building. It seems like that items above would make a good place to start: protocol Type, service and flag. There's enough variation between these that we should be able to get some base level of identification. We're also going to throw in some basic numeric data: duration, src Bytes, dst Bytes. All of these are going to be readily available from modern network equipment and should tell us a lot about what is happening on our network.

1) Feature Selection: Feature selection [3][6] is a process of selecting a subset of relevant features by applying certain evaluation criteria. In general, feature selection process consists of three phases. It starts with selecting a subset of original features and evaluating each features worth in the subset.

2) Feature Extraction: We try to form artificial features, drop features and select features as per our convenience to increase performance and accuarcy of the model.
We dropped one of the columns (Feature) (numOutboundCmds) during this process, as it was having redundant value across all the rows in the dataset. We also applied feature selection package from sckit learn to find important features, so that we don' have to use all the features to train the model.

D. TRAIN-TEST SPLIT

The Dataset provided consist of Train and test dataset separately. For training and testing, the split size is of 0.3. The prediction is being made upon 2 features, attack identification and other is of which type of attack it is.

E. MODEL TRAINING AND TESTING

We trained our model on 3 classifiers: 1. Random Forest 2. KNN 3. Logistic Regression
In Training the model, we use the split that we made and provide these split values to our classifiers. Using the train split, the model is gets trained and finally test split used to predict the outcome from the model.

Based on the nature of the data we saw above, decision trees are a good starting point for building out predictive models. In this case we'll use a random forest to build and combine multiple trees.

A confusion matrix is formed of the result to get the figure of False positives. Finally accuracy is being checked. If the accuracy is low, hyperparameter tuning is performed to amplify the accuracy and is retrained on the splits.

VII. RESULT AND ACCURACY

In this study, features from NSL-KDD were chosen using a feature selection strategy to reduce the data dimension depending on which model was trained/tested. Random feature selection is a good approach to cut down on training time and model complexity. This strategy worked well in the current model, but it may be negative in specific cases, according to the data.

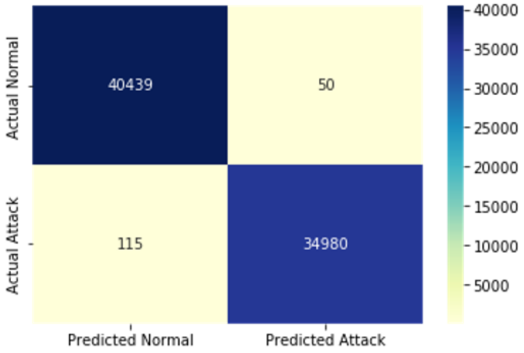


Fig. 6. The Figure shows the confusion matrix of Binary classification on Training Dataset.

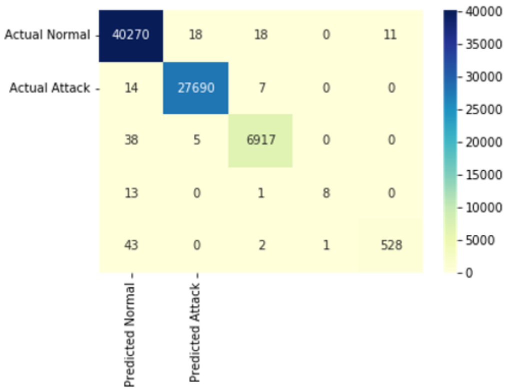


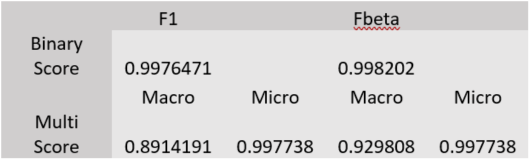
Fig. 7. The Figure shows the confusion matrix of Multi- Class classification on Training Dataset.

Random forest was used to attain an overall accuracy of above 99 percent since it can tolerate irregularly dispersed data. A random forest can detect data imbalance and uses the bootstrap process to boost minority class occurrence, minimising data misclassification and improving accuracy.

The results for all 3 classifications are shown below:

A. Random Forest Classfier

Random Forest Classifier has the maximum prediction percentage among all classifiers of 99.76% in F1 Score and 99.8% in F-Beta score.The Multi Score prediction through the classifier comes out be 89.14% (Macro) and 99.77% in (Micro) in F1 Score and 92.98% (Macro) and 99.77% in F-Beta respectively.

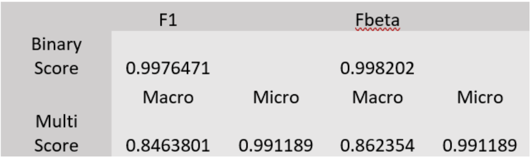


	F1		Fbeta	
Binary Score	0.9976471		0.998202	
	Macro	Micro	Macro	Micro
Multi Score	0.8914191	0.997738	0.929808	0.997738

Fig. 8. The Figure shows Binary and Multi class classification scores in Random Forest classifier.

B. KNN

KNN has the prediction percentage of 99.76% in F1 Score and 99.8% in F-Beta score.The Multi Score prediction through the classifier comes out be 84.63% (Macro) and 99.11% in (Micro) in F1 Score and 86.23% (Macro) and 99.11% in F-Beta respectively.

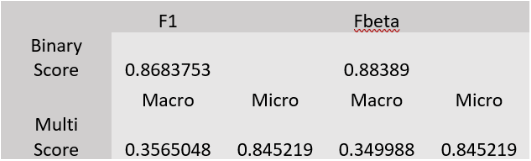


	F1		Fbeta	
Binary Score	0.9976471		0.998202	
	Macro	Micro	Macro	Micro
Multi Score	0.8463801	0.991189	0.862354	0.991189

Fig. 9. The Figure shows Binary and Multi class classification scores in KNN

C. Logistic Regression

Logistic Regression has prediction percentage of 86.83% in F1 Score and 88.389% in F-Beta score.The Multi Score prediction through the classifier comes out be and 84.52% in (Micro) in F1 Score and 84.52% (Micro) in F-Beta respectively.



	F1		Fbeta	
Binary Score	0.8683753		0.88389	
	Macro	Micro	Macro	Micro
Multi Score	0.3565048	0.845219	0.349988	0.845219

Fig. 10. The Figure shows Binary and Multi class classification scores in Logistic Regression

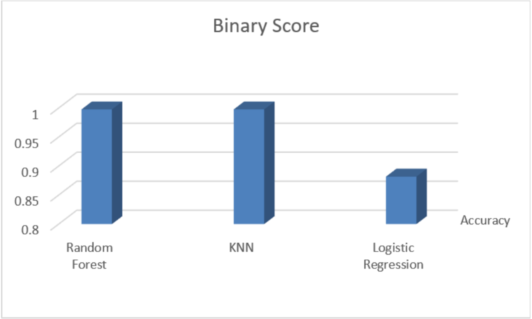


Fig. 11. The Fig show Bar Graph of Binary Score of All 3 Classifiers used in the project.

VIII. Conclusion and Future Work

In this project, we were able to explore the NSL-KDD Dataset and test the dataset on 3 Classifiers Random Forest, KNN and Logistics Regression.[9] The Training accuracy are quite appreciable and are par with the research work done on the Dataset. The findings are well discussed in the result section.

Because of their use in the NSL-KDD dataset, significant classifiers produce promising results for the Train set, but the least accurate results for the Test set.

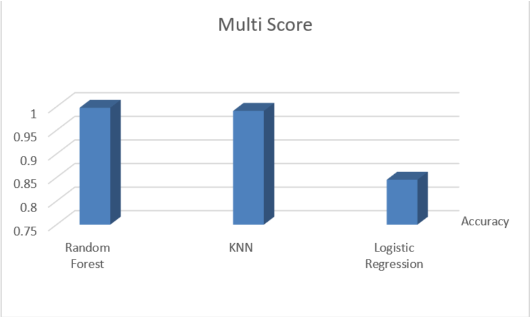


Fig. 12. The Fig show Bar Graph of Multi Score of All 3 Classifiers used in the project.

A. Future Scope

Currently, the model lags behind the test dataset, while having a high accuracy score in the training dataset. The training dataset achieved a correct prediction accuracy of nearly 99.8, but the test accuracy is still around 70. This is because there are outliers in the test dataset. In the future, it is proposed to use optimization techniques to investigate the possibility of developing more accurate intrusion detection model.

In the future, study could be focused on the possibility of using optimization techniques to provide improved IDS. As a result, ensemble-based techniques might be investigated, in which the output of several algorithms is combined to forecast the ultimate results.

Also, Using the Other classifier and techniques which were not applied during the course of this project such as Deep Learning like ANN, CNN, K-Means, RNN and to analyze their results to better select the appropriate model. Lastly the model needs to perform well on the Test Dataset which will finally make this model a success.

B. Conclusion

Because of increased unwanted access and exploitation of network resources, security has become a major problem. As attacks have become a serious problem, it is critical to detect them quickly in order to limit the damage to the system network. Machine learning classifiers have recently been popular in IDS because of their versatility, generalisation ability, and robustness. Such type of Model would be implemented on the Network (Servers) will be more robust and accurate determining the attack on the System.

References

[1] S. Aljawarneh, M. Aldwairi, and M. B. Yassein. Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, 25:152–160, 2018.

[2] A. L. Buczak and E. Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, 18(2):1153–1176, 2015.

[3] D. H. Deshmukh, T. Ghorpade, and P. Padiya. Improving classification using preprocessing and machine learning algorithms on nsl-kdd dataset. In *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, pages 1–6. IEEE, 2015.

[4] L. Dhanabal and S. Shantharajah. A study on nsl-kdd dataset for intrusion detection system based on classification algorithms. *International journal of advanced research in computer and communication engineering*, 4(6):446–452, 2015.

[5] B. Ingre and A. Yadav. Performance analysis of nsl-kdd dataset using ann. In *2015 international conference on signal processing and communication engineering systems*, pages 92–96. IEEE, 2015.

[6] H. A. Nguyen and D. Choi. Application of data mining to network intrusion detection: classifier selection model. In *Asia-Pacific Network Operations and Management Symposium*, pages 399–408. Springer, 2008.

[7] M. S. Pervez and D. M. Farid. Feature selection and intrusion classification in nsl-kdd cup 99 dataset employing svms. In *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, pages 1–6. IEEE, 2014.

[8] R. D. Ravipati and M. Abualkibash. Intrusion detection system classification using different machine learning algorithms on kdd-99 and nsl-kdd datasets-a review paper. *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, 11, 2019.

[9] S. Revathi and A. Malathi. A detailed analysis on nsl-kdd dataset using various machine learning techniques for intrusion detection. *International Journal of Engineering Research & Technology (IJERT)*, 2(12):1848–1853, 2013.

[10] S. Solanki, C. Gupta, and K. Rai. A survey on machine learning based intrusion detection system on nsl-kdd dataset. *Int. J. Comput. Appl*, 176:36–39, 2020.