International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

AN INTRUSION DETECTION FRAMEWORK BASED ON BINARY CLASSIFIERS OPTIMIZED BY GENETIC ALGORITHM

Aswathy T[1], Misha Ravi[2]

[1]*M. Tech. Student, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.*
[2]*Assistant Professor, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.*

Abstract - *Intrusion detection system is one of the key aspects of security in today's network environment. The systems deal with attacks by collecting information from a variety of system and network sources and then identifying the symptoms of security problems. This work analyzes four machine learning algorithms such as Decision Tree, Naive Bayes, Support Vector Machine and K-Nearest Neighbour to detect intrusions with three performance criteria: precision value, accuracy, recall. The purpose of this system is to find out which classifier is better in its intrusion detection system process. The primary target is to build up an application that can process incoming network connection and discover the hazard of intrusion. The framework that aggregates distinct classifiers and discover the normal and abnormal network connections. A genetic algorithm is used to generate massive-quality solutions from a set of classifiers. As the experimental results, SVM gives better results and it efficiently detects intrusion with an accuracy of 99%. Experimental results show that the proposed method outperforms baselines with respect to various evaluation criteria.*

Key Words: **Intrusion Detection, NSL-KDD, Binary Classifiers, KNN, Genetic Algorithm.**

1. INTRODUCTION

Network attack detection is one of the most important problem in network information security. The security of the network is any activity designed to safeguard network connection and data's usability and integrity. It is the responsibility of effective network security to manage network access. Security of the network includes policies and practices. It is used to avoid and inspect malicious software, modification or a denial of a computer network and network resources that are accessible. Network security involves authorizing data access in a network controlled by the administrator of the network. Security is a primary issue in all types of networks, especially in large organizations. In today's world, protecting computer resources and stored documents is a major concern. Current security solutions that are secure organizations but they do not discover all kinds of attacks. More security mechanisms are needed, like intrusion detection systems, because firewalls are unable to detect network attacks because they are mostly deployed at the network boundary and thus only control traffic that enters or leaves the network. There can be a huge percentage of intrusions in the network and intrusion can monitor and analyse different network events and, if the system has been misused, report to the administrator immediately. The intrusion detection system's purpose is to assist computer systems to manage attacks. The hybrid intrusion detection system is a powerful system. However, by combining multiple techniques into a single hybrid system, an intrusion detection system can be created that has the advantages of multiple approaches while overcoming many of the disadvantages. While it is true that combining multiple different intrusion detection system can generate a much stronger intrusion detection system theoretically. Various intrusion detection system technologies examine traffic and look in different ways for intrusive activity. It can be a very challenging task to get multiple intrusion detection system approaches to coexist in a single system.

This work focuses on the intrusion detection based on data mining that is optimized by genetic algorithm. Data mining based intrusion detection systems are considered and analysed for the efficient attack detection purpose. The analysis and survey of intrusion detection system will be brought into focus by examining several case studies and innovative solutions. Hybrid approaches have become the mainstream in intrusion detection system studies as they are superior in terms of accuracy to the single classification technique. The project's main objective is to develop an application capable of processing incoming network connection and identifying the intrusion risk.

1.1 Problem Statement

An intrusion detection system is a network security countermeasure that monitors a network for suspicious activity. The main objective of the project is to develop an application that can process incoming network connection and identify the risk of intrusion. The framework that aggregates different classifiers and find the normal and abnormal network connections and a genetic algorithm is used to generate high quality solutions from a set of classifiers.

2. METHODOLOGY

The methodology describes the existing and proposed system.

2.1 Existing System

Network intrusion detection system is a software that detects abnormal system usage by monitoring and analyzing the network environment. It is important for organizations and individuals to secure computer and network information because compromised information can cause considerable damage. In order to avoid such circumstances, intrusion detection systems are important. In the past few decades, many intrusion detection systems have been developed. Most of the intrusion detection systems are misuse detection and anomaly detection. Intruders with known patterns are detected by misuse detection systems. Anomaly detection systems detect deviations from normal behavior in the network. Such intrusion detection systems have limitations. Therefore, a hybrid system is needed. However, a system of intrusion detection can be created by combining multiple techniques into a single hybrid system, which has the advantages of multiple approaches while overcoming many of the disadvantages.

2.1.1 Drawbacks of Existing System

Intrusion detection become an increasingly important technology that monitors network traffic and identifies network intrusions. In the detection of intrusion, different approaches are used. This intrusion detection has drawbacks. Drawbacks of existing system are most existing systems for intrusion detection only determine the occurrence of attacks, but do not provide their type and also have low detection performance for low frequency attacks [1]. The data set for intrusion detection is imbalanced. Low - frequency attacks have few instances compared to high - frequency attacks and can be considered as outliers. Accurate information on intrusion is very important for network administrators to take appropriate security measures. Another limitation is too many parameters in certain intrusion detection system. That is, some intrusion detection models, have many parameters. Setting values for those parameters is not easy. The system therefore needs an optimization algorithm. Unoptimized values can have a negative impact on detection performance. Due to these limitations binary classification based intrusion detection system is proposed.

2.2 Proposed System

An intrusion detection system is a security mechanism to minimize risk of intrusion. The identification of an intrusion type is more valuable than merely determining that an attack occurred to protect network security. It is essential to provide network administrators with the exact intrusion information in order to take appropriate action to secure the computing infrastructure. Different techniques are being used to detect intrusion, but their performance is an issue. The performance of intrusion detection depends on the accuracy and precision needed to improve the detection rate. An efficient classification is needed to resolve performance concerns. To address the performance and accuracy challenges associated with detecting intrusion, this work attempts to prove effective in addressing the classification problem.

The title of the proposed system is "An intrusion detection framework based on binary classifiers optimized by genetic algorithm." It is a hybrid system for detecting intrusion and is designed based on binary classification and K-Nearest Neighbor algorithm. The data mining framework has been experimentally proven to detect intrusion based on binary classification. Decision tree, Naïve Bayes and Support Vector machine (SVM) act as binary classifiers. The input data of the network connection is converted to dataset and submitted to different classifiers in order to classify connections as normal or attack. In data capturing, firstly the packets are captured. Wireshark is used for packet capturing. Then the dataset is upload and it is stored in a data table. The dataset used is NSL-KDD and it contains 41 attributes. These 41 attributes are classified into four different categories such as Basic, Content, Traffic and Host. Using these data, the CSV (Comma - Separated Values) file is created and is converted into ARFF (Attribute-Relation File Format) file. Firstly, decision tree, then Naïve Bayes and last support vector algorithms are applied to the ARFF file. Precision value, recall, accuracy and performance of the algorithms are determined. Using these algorithms, to identify the network connection is normal or attack. Then the probability of intrusion is identified in the decision tree, naïve Bayes and support vector classifiers. Aggregate these probabilities and create a CSV file using these aggregation results. Also, the CSV file is converted to ARFF file. The generated ARFF file is fed to a K-Nearest Neighbor (KNN) classifier and predict the normal connection and attack. The proposed system architecture is shown in Fig -1.

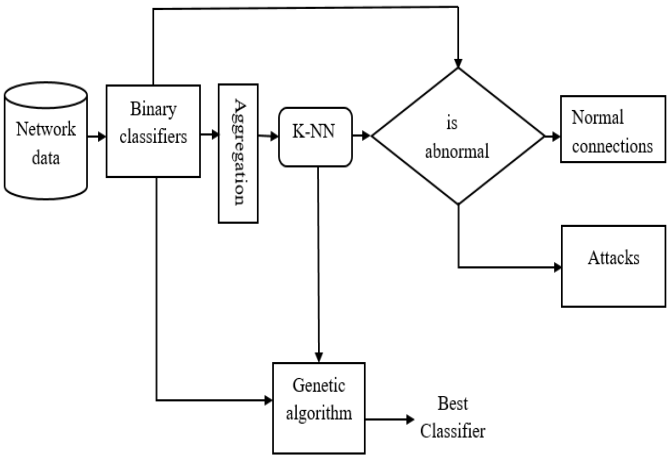

Fig -1: Proposed System Architecture

The test set will be created based on the four categories such as basic, content, traffic and host. Using the test set, the

decision tree, naïve bayes, support vector machine and k-nearest neighbor algorithms are implemented. These algorithms are used to determine precision, recall and accuracy with test set values. Then determine the network connection is normal or attack by using the test set data. A genetic algorithm is used to optimize these algorithms, which means that the genetic algorithm gives better results in intrusion detection and also calculate the optimized algorithm's best fitness value. Upon detection of an intrusion, proactive steps are taken to generate an SMS alert. So the administrator can take relevant action.

3. SYSTEM DESIGN

3.1 Modules and Their Functionalities

Network detection plays a very important role in protecting the security of computer networks. Intrusion detection based on a binary classification is experimentally proved by implementing a data mining based framework. The input data arrived from network connections given to binary classifiers and KNN algorithm to classify the connections as normal or attack. The main part of the proposed system is the administrator module. The administrator controls the whole process of the system The intrusion detection system consists of seven modules:

1. Packet capturing module
2. Dataset management module
3. Class detection module
4. Prediction module
5. Genetic algorithm based prediction module
6. Evaluation module
7. Intrusion prevention module

3.1.1 Packet Capturing

A live network capturing section is implemented in packet capturing module. Wireshark is used to capture network traffic. The packet information stored in a data table. This module is not completely implemented. In this proposed system, any hardware devices do not used for packet capturing. So, limited parameters are obtained in this module. The original dataset contains many parameters. The main goal of packet capturing module is to implement a real time intrusion detection system. Using the limited parameters, implementation of a real time intrusion detection system is not possible.

3.1.2 Dataset Management

In dataset management module, the NSL-KDD dataset is used. The dataset contains 41 attributes and a 'class' attribute. The class attribute indicates whether a given instance is a normal connection instance or an attack. In this step, the datasets with different size taken from NSL-KDD are input to the framework. These 41 attributes can be divided into four different categories like Basic, Content, Traffic, Host [2].

- Basic (B) features are individual transmission control protocol connection attributes.
- Content (C) parameters are the attributes suggested by the domain knowledge within the connection.
- Traffic (T) features are the calculated attributes of the two-second time window.
- Host (H) features are attributes designed to evaluate more than two seconds of attack.

The classified attributes are stored in a data table. It will be used as the input of the proposed intrusion detection system.

3.1.3 Class Detection

In class detection module, test sets are containing incoming connections that does not have a class label. Different classes can be mapped to an incoming connection by the prediction algorithm. The original dataset is used as training set. The dataset consists of normal connection and attack type. Neptune, ipsweep, nmap, port sweep, Satan, teardrop, smurf are the attack types. The original dataset is partitioned into four different categories and using this data to create a CSV (Comma - Separated Values) file. A file of comma - separated values is a text file that uses a comma to separate values. A CSV file stores plain text tabular data. Each file line is a record of data. Each record, separated by commas, consists of one or more fields. The generated CSV file is converted to ARFF (Attribute-Relation File Format) file. An ARFF file is an ASCII text file that shows a list of instances that share a set of attributes. This ARFF file is the input of different classifiers.

3.1.4 Prediction

The prediction module decides whether a network connection is attack or normal. It can be done by using, Decision tree, Naïve Bayes, Support Vector machine and K-Nearest Neighbor algorithms. Input of these predictive algorithms is a CSV file. The file in csv contains network data.

(i) Decision Tree

Decision tree is the strongest and most popular classification and prediction tool. A decision tree is a tree- like flowchart in which each internal node denotes a test on an attribute, each branch represents a test results, and each leaf node has a class label. It is a supervised classical learning algorithm. The main concept behind the learning of the decision tree is the following: to build a predictive model that is mapped to a tree structure starting from training data. The C4.5 algorithms is implemented in Weka as a classifier called J48 to build decision trees. Classifiers are organized in a hierarchy like filters: J48 is called in its full name weka. classifiers. tree. J48. Furthermore, in training sample set of abnormality and noise will also cause some abnormal branches, so the decision tree will need to be pruned. In this algorithm, the decision is followed from the root to a leaf to

classify an item [3]. An attribute is tested at each node and the corresponding branch is followed based on the result. This procedure goes on until a leaf has been reached. The J48 algorithm classifies an unknown network connection as an attack or normal data through this model.
The advantages of Decision tree are:

1. Simple to understand and interpret.
2. A variety of input data such as Nominal, Numeric and Textual can be handled.

(ii) Naïve Bayes

Naive Bayes is a probabilistic algorithm family that uses the probability theory and the Bayes theory to predict the sample category. It uses a simple implementation of the Bayes Theorem where the preceding probability for each class is calculated from the training data and assumed to be independent. A probabilistic classifier is a classifier that can predict a distribution of probability over a set of classes, given a sample input. Based on the dataset provided with training data, it is used to categorize the results for any test data. Class for a classification of Naive Bayes using classes of estimators. Precision numeric estimator values are selected on the basis of training data analysis. Naive Bayes estimates prior probability for each class and makes the class's highest probability prediction. As such, it supports both binary and multi-class classification issues. It makes the assumption that the effect on a given class of a variable value is independent of other variables values. This model classifies an unknown network connection as an attack or normal data in the network.

The advantages of Naïve Bayes are:

1. Easy to understand and implement.
2. To estimate the parameters, it requires a small amount of training data.

(iii) Support Vector Machines (SVM)

The support vector machines are known as the best binary classification learning algorithm. The training algorithm for support vector machine sets up a model that assigns a new occurrence to a class or other class, making it a probabilistic binary linear classifier for SVM. Recently, intrusion detection has also been applied to information security [4]. Because, of its good generalization character, support vector machine has become most popular intrusion detection techniques. Support vector machines are supervised models of learning with associated learning algorithms that examine data for analysis of classification and regression.

Support vector machines are based on the idea of planes of choice that define boundaries of decision. A decision plane is a decision plane that separates a group of objects with different class memberships. By separating the input space from linear or nonlinear surfaces, classification is achieved. The separation can be expressed as a linear kernel combination in support vector classification. Support vector algorithms can converge to accurate solutions and calculate an error effectively. By adding one point at a time to the current set of data, the algorithm works. The support vector machine model is designed and trained using a reduced NSL-KDD dataset. It uses a high dimensional space to find a binary classification hyper plane where the rate of error is minimal. By this model, the support vector machine will classify an unknown network connection as an attack or normal.

The advantages of SVM are:

1. SVM provides a good generalization nature.
2. Flexibility
3. It has the ability to update the training patterns dynamically.
4. High accuracy

(iii) K-Nearest Neighbor algorithm (KNN)

K-Nearest Neighbors is one of Machine Learning's most basic but essential classification algorithms. It is a part of supervised learning and in pattern recognition, data mining, and intrusion detection, it finds intense application. This algorithm is intended to classify a new object based on attributes and samples of training. The k-nearest neighbor algorithm in Weka is called IBk (Instance Based Learner). The IBk algorithm does not create a model, but a prediction for a test instance is generated. For each test instance, the IBk algorithm uses a distance measure to locate k "close" instances in the training data and to make a prediction using those selected instances. It estimates the distance between the various input vector data points and assigns the unlabeled data point to its closest neighbor class. K is a key parameter. If value of K is large, it will take a long time to predict and influence the accuracy by reducing the noise effect. Here, the probability of intrusion is identified in the decision tree, naïve bayes and support vector machine classifiers. Then aggregate these probabilities and create a CSV file using these aggregation results. Also, the CSV file is converted to ARFF file. The generated ARFF file is fed to a KNN classifier and predict the normal connection and attack type.

The advantages of KNN are:

1. Robust to noisy training data.
2. Effective if the training set is large.
3. Easy to implement for multi-class problem.

3.1.5 Evaluation

Intrusion detection using different strategies will be evaluated by analysing the precision, recall, accuracy and performance comparisons also made with different datasets.

For the purpose of this project, precision value, recall, accuracy is used as evaluation criteria.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/TP+FP

Recall = TP/TP+FN

Where TP is the True positive, which are actually an attack and classified as abnormal. TN is the True negative, it represents the number of normal connections correctly classified as normal. FP is the False positive which is actually normal, but classified as an attack and FN is the False negative, which is the number of attacks incorrectly classified as normal.

3.1.6 Genetic Algorithm Based Prediction

A genetic algorithm is an optimization technique. Genetic algorithm works with optimal cross-value based on the best solution. It can address a variety of optimization techniques provided they can be parameterized in such a way that a solution to the problem provides a measure of how accurate the algorithm-finding solution is. This measure is defined as fitness. Different algorithms tested in the existing system are fed into a framework of genetic algorithms. The best scenario of prediction will be generated as output. Support vector machine classifier is obtained as the best classifier for intrusion detection in terms of accuracy using genetic algorithm.

3.1.7 Intrusion Prevention

Generally, when an intrusion is detected by the admin, proactive steps are taken as an SMS alert will be generated. This work is based on data set, not real-time data, so prevention method is possible to prevent intrusion only by SMS alert. If real data is used, the self-shutdown utility will be activated. Using test data, to detect incoming network connection is normal or attack. If it is a type of attack, an SMS alert will be generated in the intrusion prevention module and sent to the administrator. The administrator can take appropriate action.

4. RESULTS AND ANALYSIS

4.1 Result

This experimental results of the proposed system an intrusion detection framework based on binary classifiers optimized by genetic algorithm is discussed in this section. The system that uses the operating system for windows 10 and windows platforms here is c#.net. And the database created is a SQL server. The proposed system is using network data for results assessment. Only a few public datasets are available in the field of intrusion detection to evaluate the performance of intrusion detection systems. An effective benchmark is the NSL-KDD dataset. Each instance includes 41 input features and a class label in the NSL-KDD dataset. The class label specifies whether an instance's status is either normal or attack.

The proposed system is implemented using admin module and different sub-modules. The system's input is the NSL-KDD dataset. The dataset can be divided into basic, content, traffic and host categories. Using this dataset CSV file is created and it is converted to ARFF file. Decision tree, Naïve Bayes, Support Vector Machine and K- Nearest Neighbor algorithms are used in this system. The ARFF file is applied to Decision tree, Naïve Bayes, Support Vector Machine algorithms and finding the network connection is normal or attack. And also, when using each algorithm, calculating precision, recall, accuracy, performance time. Then calculating the average probability of the possibility of intrusion in each algorithm. The calculated average probabilities are aggregated and the aggregated result is used for the input of KNN algorithm. This algorithm also detecting the normal connection and attack. Using test set, identify the network connection is normal or attack. And also, calculating precision, recall, accuracy, performance time. Then genetic algorithm is applied to the result of test set and find the best classifier on basis of precision and recall.

4.2 Analysis

This system shows intrusion detection based on binary classifiers. It also shows analysis based on algorithms and evaluation criteria which are used in the system.

4.2.1 Analysis: Precision, recall, accuracy with algorithms for Basic dataset category.

This analysis shows three evaluation criteria such as precision, recall, accuracy with different algorithms like decision tree, naïve bayes, support vector machine and k-nearest neighbour. The X- axis shows values like precision, recall and accuracy. Y- axis is the various algorithms such as decision tree, naïve bayes, svm and knn. Fig. 2 shows performance evaluation of algorithms with basic dataset.

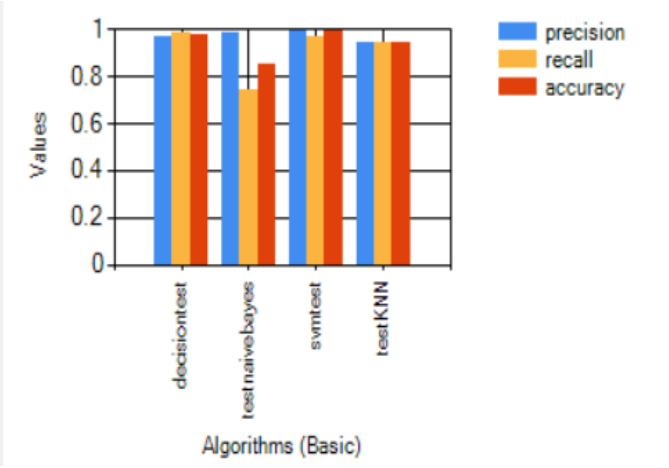

Fig -2: Analysis of precision, recall, accuracy values with algorithms for Basic dataset.

4.2.2 Analysis: Precision, recall, accuracy with algorithms for Content dataset category

Fig. 3 shows performance evaluation of algorithms with content dataset.

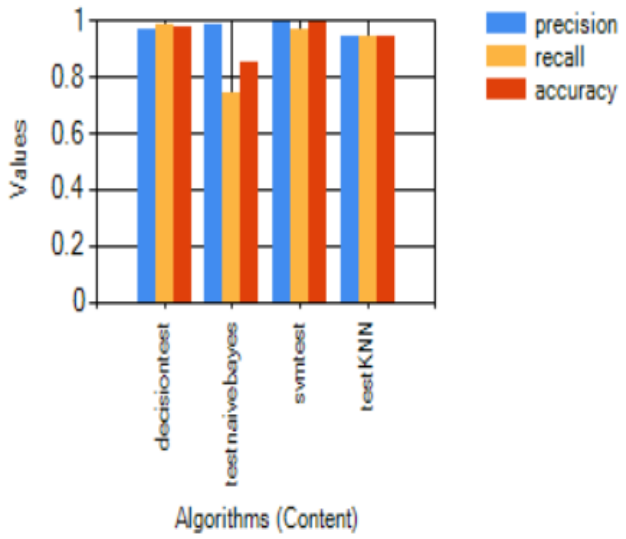

Fig -3: Analysis of precision, recall, accuracy values with algorithms for content dataset.

4.2.3 Analysis: Precision, recall, accuracy with algorithms for Traffic dataset category

Fig. 4 shows performance evaluation of algorithms with Traffic dataset.

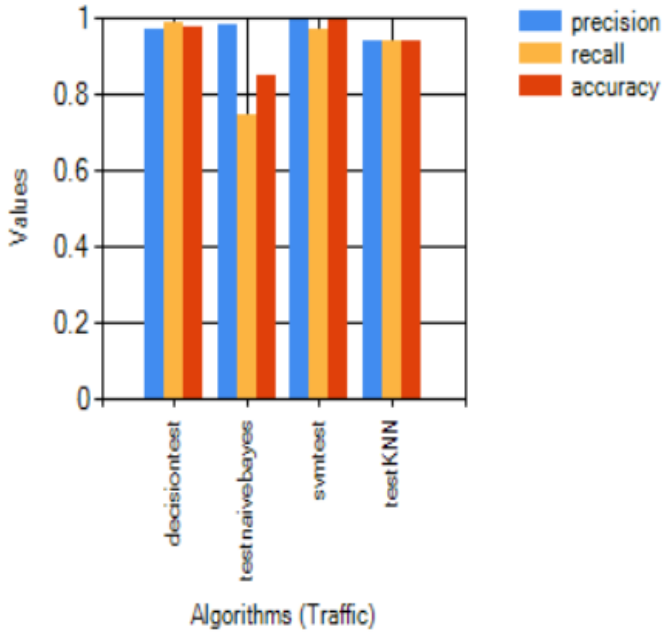

Fig -4: Analysis of precision, recall, accuracy values with algorithms for Traffic dataset.

4.2.4 Analysis: Precision, recall, accuracy with algorithms for Host dataset category

Fig .5 shows performance evaluation of algorithms with Host dataset. X- axis shows various algorithms like decision tree, naïve bayes, svm and knn. The Y-axis shows precision, recall and accuracy values.

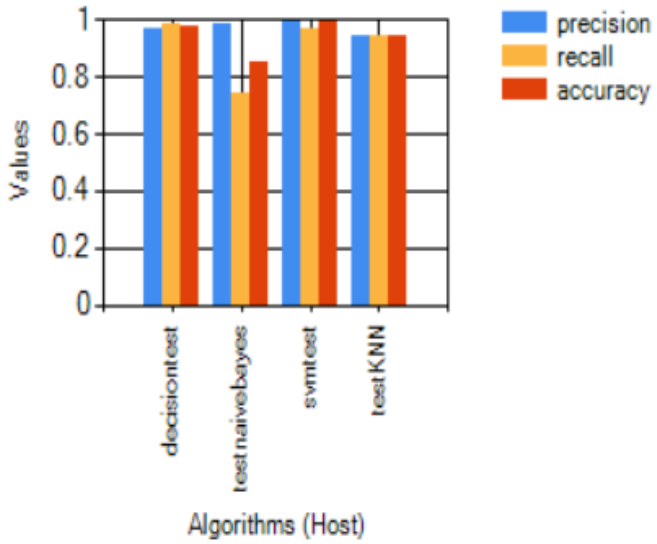

Fig -5: Analysis of precision, recall, accuracy values with algorithms for Host dataset.

5. CONCLUSION AND FUTURE WORK

The proposed system is an intrusion detection system based on binary classifiers and a genetic algorithm. The goal of this system is to develop an application designed to process incoming network connection and identifying the intrusion risk. The framework that aggregates different classifiers and finds normal and abnormal network connections, and a genetic algorithm is used to generate high-quality solutions from a set of classifiers. In this system, the performance measurement of four machine learning classifiers such as Decision tree, Naive Bayes, Support Vector Machine and K-Nearest Neighbor is compared at the task of detecting intrusions and found that Support Vector Machine is excellent in performance in terms of accuracy compared to other classifiers. The four algorithms are suitable for detecting intrusions from the NSL-KDD dataset. Support Vector Machine was found to be much better at detecting intrusions with an accuracy rate of about 99 percent than Decision tree, Naive Bayes and K-Nearest Neighbor classifiers.

This also opens up new possibilities for future work, including: to use real-time network data to detect intrusion and also to introduce classification methods to improve the precision, recall and accuracy of intrusion detection.

REFERENCES

[1] L. Li, Y. Yu, S. Bai, Y. Hou and X. Chen (2018) An Effective Two-Step Intrusion Detection Approach Based on Binary Classification and k -NN, in *IEEE Access*, vol. 6, pp. 12060-12073.

[2] Aggarwal, Preeti & Sharma, Sudhir. (2015). Analysis of KDD Dataset Attributes - Class wise for Intrusion Detection. Procedia Computer Science. 57. 842-851. 10.1016/j.procs.2015.07.490.

[3] M. Kumar, M. Hanumanthappa and T. V. S. Kumar, "Intrusion Detection System using decision tree algorithm," *2012 IEEE 14th International Conference on Communication Technology*, Chengdu, 2012, pp. 629-634.

[4] Adhi Tama, Bayu & Rhee, Kyung Hyune. (2017). Performance evaluation of intrusion detection system using classifier ensembles. International Journal of Internet Protocol Technology. 10. 22. 10.1504/IJIPT.2017.083033.

BIOGRAPHIES

Aswathy T, she is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Elavumthitta, Kerala, India. Her research area of interest includes the field of data mining, internet security and technologies

Misha Ravi received the master's degree in Software Engineering from Cochin University of Science and Technology, Kerala. She is an Assistant Professor in Department of Computer Science and Engineering, at Sree Buddha College of Engineering.