INTRUSION DETECTION SYSTEM USING HYBRID ENSEMBLE MACHINE LEARNING TECHNIOUES

AJEESHA M I
Ph.D Research Scholar
School of Computer Studies
Rathnavel Subramaniam College of Arts and Science
Coimbatore, Tamilnadu, India,
ajeesha1393@gmail.com

DR. D FRANCIS XAVIER CHRISTOPHER
Director, School of Computer Studies
Rathnavel Subramaniam College of Arts and Science
Coimbatore, Tamilnadu, India,
christopherd@rvsgroup.com

***Abstract*— Global communication is easier today with digitization. The internet became widespread in less than a vicennial. Machine learning makes todays life easier. Without human intervention system doeseverything. In this paper an ensemble approach of coalesce the machine learning algorithms together by analyzing the performance metrics is done. The dataset used for the model is kdd cup 99. The two partsof the dataset includes: KDD-Train and KDD-Test. The performance metrics are evaluated using the ensemble model. The classifiers used for the experiment is Naive Bayes Classifier, Decision Tree Classifier and Logistic Regression. The hybrid ensemble model is a predictive method that cooperate the outcome from two or more models. The results of the analysis are discussed in terms of performance metrices. The ensemble model is then compared with individual machine learning models. The improved performance of ensembles often reduces the variance component of the prediction inconsistencies made by the models.**

Keywords — *Supervised machine learning, Naïve bayes classifier, Logistic regression, Decision tree classifier, Kdd cup 99.*

I. INTRODUCTION

Machine learning emerges as a technological powerhouse. It is dominating day by day. Machine learning allows a system to learn the data provided toit. Machines will analyze abounding of data quickly and efficiently than human. Machine learning come under the broad field of artificial intelligence and leads to the system to learn from data, observe patterns and take decisions without manual intervention. Machine learning is required for this innovative world. Along with artificial intelligence (AI) the machine learning represents a significant step in digitization.

The machine learning algorithms use recorded data as input to predict new results. The learning process begins with observations and results the predictionsdetermined from the given data provided. The objective is to perform the system to learn without human assistance and modify actions accordingly.

Machine learning is a developing trend in the technological world. Network security is required for all internet users. The IDS uses machine learning to build a model maintaining normalactivity and then analyze new behaviour with the existing model.

Network intrusion is an apocryphal activity on the computer network. Intrusion Detection System reveals the network intrusion using numerous machine learning algorithms. It builds a predictive model efficient to distinguish between bad connections(intrusion/attacks) and good (normal) connection. IDS can accomplish to monitor networktraffic of a specific device or to monitor all networkscommonly used.

The KDDCup99 dataset is officially recognized in academic research. It is used in many IDS and machine learning studies. It includes 41 features andis arranged as either normal or an attack of any specific model. The attack falls into any of the four attack types such as: Denial of service attack (DoS),User to root attack(U2R), Remote to local attack(R2L) and Probing attack. The dataset contains 24 training attack types and further 14 types ofattacks include in the test dataset.

The IDS are of two types to detect attacks: anomalybase intrusion system and misuse base intrusion system. Anomaly intrusion detection system perform to detect attacks in accordance with the recorded normal behavior. A behavior-based anomaly is a newer technology is designed to monitor unknown attack. It detects previously unknown attacks thus leads to false positives. Misuse intrusion detection system can assuredly detect reveal attacks and is unobtainable to find newattack. Thus, it has no false alarm [1].

IDS identifies all traffic flow both in incoming and outgoing directions. They scan and examine the type of payload or content of the packets. A confusion matrix is used to find out the performance of a model, and has been able to classify the data points belonging to each classes. The performance metrics accuracy, precision, recall and weighted f1- score are calculated with respect to the confusion matrix to determine the best model.

Ensemble learning provides better predictive performance by integrating the predictions from multiple algorithms. In this machine learning approach by which multiple models are combined to solve a particular computational problem to improvethe prediction. An ensemble model combines the predictions of multiple models. The models are referred as ensemble members may be the same typeor different types. The predictions are made by usingstatistics, the mean or mode, or by more sophisticatedmethods. An ensemble model results

well predictions than a single predictive model and reduces the dispersion of the results [2].

The remaining section is ordered as follows: Section 2 explains the related work used in this study. Section 3 provides the hybrid model evaluated for this work and the classification methods. The proposed modelling of the research work is explained with evaluation metrics and the datasets is illustrated. Section 4 includes results and discussions of the research work are provided. Finally, Section 5 provides the conclusion of theresearch work.

II. RELATED WORK

Zina Chkirbene and Sohaila Eltanbouly suggest a hybrid approach of integrating two machine learning algorithms to detect the different possible attacks byperforming effective feature selection andclassification. This system uses Random Forest algorithm for the feature selection. Theclassification method called Classification andRegression Trees (CART) is used. The dataset usedhere is UNSW-NB15 results the proposed method with a good performance compared to the existing algorithms [3].

Muhammad Ashfaq Khan and Yangwoo Kim [4] develop a hybrid intelligent intrusion detection system (HIIDS) to study the significant features representation reliably and mechanically from enormous unlabelled raw network. The dataset usedis ISCX-UNB. To evaluate the efficiency of a model, LSTM is used to detect temporal features and an AE to more efficiently detect global features. The experimental results shows a better accuracy rate of97.52% demonstrating the spark MLlib and LSTMAE-based HIIDS in 10-fold cross-validation test.

Harsh H. Patel and Purvi Prajapati experimented the decision tree exceeds other classifiers with respect to accuracy, time and precision [5]. Hatim Mohamad Tahir and Wael Hasan implemented a hybrid intelligent method for network intrusion detection. They combine K-Means clustering and SVM classifier. The goal of this work is to minimize the false positive and false negative values results to make better detection rate using NSL-KDD dataset. Using support vector machine, the classification process is done [6]. The result work achieved good detection rate and lower false alarm rate after training and testing the hybrid algorithm. Hind Bangui, Mouzhi Ge [7] implemented a new machine learning model. Random Forest and a posterior detection with coresets are used to improve the accuracy and increase efficiency. The proposed machine learning model can obtain better detection accuracy compared to other machine learning models.

Po-Jen Chuang and Si-Han Li proposed a hybrid model by integrating two algorithms Naive Bayes and C4.5. The performance of trained classification model is improved according to the training time of network traffic. Comparing with other hybrid machine learning algorithms, the developed model is able to shorten the needed training time and yields desirable detection performance [8].

Tuan A Tang and Lotfi Mhamdi [9] develop a Deep Neural Network (DNN) model to detect an intrusion. The NSL-KDD Dataset train the model. The six basic features of NSL-KDD Dataset are evaluated. The results show strong potential for flow-based anomaly detection in SDN.

Unal Cavusoglu suggests a hybrid and layered model for Intrusion Detection System (IDS). Combining different machine learning algorithms with feature selection techniques is used in different attack types for high performance intrusion detection. The data pre-processing is done with the dataset NSL-KDD. Different feature selection algorithms reduce the size of the dataset. Two new methods have been proposed for feature selection method. The layered architecture is designed by analysing suitable algorithms based on attack type. To demonstrate the outcome of the proposed system, it is compared with the other methods and performance evaluation is done. It results the proposed model has high accuracy and a low false positive rate in all attack types [10].

III. PROPOSED MODELLING

3.1 Data Mining

Data mining is designed as a superset of many distinct methods to extricate insights from data. It requires traditional statistical methods and machine learning. Data mining interrelates techniques in varied fields to verify the unknown patterns from data. Data mining also comprises of the research of data storage and data utilization. Data mining deal with already existing patterns in the data, machine learning predicts the outcomes with respect to the results obtained in the past using the pre-existing data. Data mining is implemented on current dataset to find patterns like a data warehouse. Machine learning, is trained on a ‘training’ data set, which instruct the system how to follow data, and then to make predictions about new data sets. One of the aims of machine learning is data mining it extracts more accurate data. This eventually helps to upgrade the machine learning to attain better results.

3.2 Machine learning

It is a systematic search of computer algorithms that can upgrade deliberately through the use of data. Machine learning come under artificial intelligence. The machine learning model test for a validation error on new data. It employs a sequential method to learn from data. The Shallow Learning algorithms involved for IDS are Decision Tree, K-Nearest Neighbor (KNN), K-Mean Clustering and other Ensemble Methods. The machine learning methods widely adopted are supervised learning and unsupervised learning and reinforcement learning. Supervised learning is a branch of machine learning. The machine is trained well on labelled data in a supervised learning method. The input has to be labelled accurately in supervised learning. The supervised learning is exceptionally dominant under the right circumstances. Unsupervised machine learning able to work with unlabelled data. Unsupervised learning is data-driven based upon the data and its properties. The outputs of an unsupervised learning task are managed by the data. Reinforcement learning is another branch of machine learning algorithm where program communicate with the conditions and produce output within that. In this method it

trains the machine learning models to make a chain of decisions.

3.3 Supervised learning

In supervised learning, the model is maintained on a specified dataset with both raw input data and its outcomes. The training dataset and test dataset are the two parts of a dataset where the training dataset trains the network model whereas the test dataset acts as new data for predicting outcomes and verify the accuracy of the model. The execution of the model isfast because the training time utilized is less as the desired results is already in the dataset. This model predicts accurate results on new data without even knowing a prior target. In some of the supervised learning models, the output result is relearned in order to achieve the highest possible accuracy [12]. Each input instance has an expected value associated with it, the value can be discretely real or continuous value. The algorithm predicts the output after training the input patterns.

3.4 Classification Method

Classification come under the supervised machine learning approach. The algorithm works based on the input data provided to the model and then classify new observations based on the input. Classification modularizes the given dataset into classes, it is suitable on structured data and unstructured data. The class also known as label is predicted first with respect to the given data points. Classification is the method of evaluating, realizing and grouping the data and objects into subcategories, it is a form of pattern recognition. Classification method use the input data to predict the output data with predetermined data. One of the applications of classification method is spam filtering. There are different classification algorithms depending on the dataset the algorithm is selected.

3.5 Types of Learners in Classification

Lazy Learners – Lazy learning is an instance-based method. They do not generalize until they needed. It can create many local approximations. Lazy learners have the input data and hold back until a testing data is created. It is slow as it is calculated based on the current dataset instead of an algorithm with historic data. It can represent more complex function. The predicting time is high with respect to the eager learners. Eg: k-nearest neighbor.

Eager Learners – Eager learners builds a classification model before seeing the query based on the given training data. It is fast, as it has precalculated algorithm. It can create global approximation. When it gets a dataset, it builds a model for the whole dataset at first.

3.6 Bagging Method

Bagging, also called as bootstrap aggregation an ensemble learning method that is mainly used toat same time. decrease the variance in a noisy dataset. In bagging, a sample of data in a training set is selected with replacement. Each data points can be taken more than once to reduce the variance value and attains higher stability with minimal errors. Bagging combines same type of predictions. Models are built independently. The weak models are trained parallel.

Bagging consists of aggregation and bootstrapping. Bootstrapping is a sampling method. The replacement method is used to select a sample out of a set. The selected samples execute the learning algorithm. The bootstrapping method uses sampling with substitutions to build a random selection. If the selection of a sample is not in replacement method, then the upcoming selections of samples dependent upon the previous selections, thus the procedure become non-random.

Boosting combines different types of predictions and the weak learners are trained sequentially. Different models are created and with each new model iteration, the weights are increased in the previous model of misclassified data. Boosting decreases variance value and also provide high stability with minimal errors. It is implied when the classifier is unstable and having high variance. The predictions of the model go through aggregation to combine them for the final prediction to verify all the possible outcomes. The aggregation can be done based on the probability of predictions derived from the bootstrapping of every model in the procedure

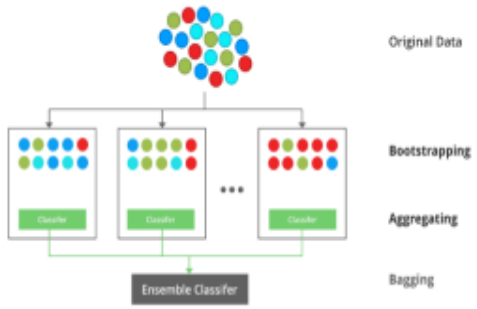


Fig. 1. Bagging Method of ensemble classifier

The figure above shows the bagging method of an ensemble classifier. Bagging influences a bootstrapping sampling technique to create distinct samples. This resampling method creates different subsets of the input dataset by selecting data points at random with replacement. That is, whenever selecting a data point from the training dataset, it can able to select the same instance multiple times. Thus, resulting a value or instance repeated twice or more in a sample. Bagging reduces the variance value within a learning algorithm. This is better with high-dimensional data, where missing values can increase higher variance, making it more prone to overfitting and preventing accurate generalization to new datasets.

3.7 Intrusion Detection System

IDS are software application that play a main role in modern organizations to protect against intrusions and malicious activities. We have two majors intrusion detection system categories:

• *Host Based Intrusion Detection Systems (HIDS):* It monitors the host devices to identify the threats inside the network.

• *Network Based Intrusion Detection Systems (NIDS)*: It detects the network anomalies by monitoring the inbound and outbound traffic.

The intrusion detection techniques are of two types:

• *Signature based detection technique*: It detect the intrusions by evaluating the traffic and compared the patterns of the signatures of known threats.

• *Anomaly-based intrusion technique*: Evaluates the normal traffic activity with respect to the networks against the baseline.

3.8 KDD Cup99 Dataset

A dataset can hold a set of information needed by applications or operating system itself. The kdd dataset is most widely used dataset for intrusion detection. It is used to build predictive models efficient for differentiating between intrusions or attacks, and normal connections. The kdd database contains a standard dataset total of 4898431 instances with 41 attributes. Each relation contains 100 bytes and is described as either normal or as an attack, with absolutely one labelled attack type. Basically, the attacks fall into four main groups:

DOS: denial-of-service attack

R2L: unauthorized access from a remote machine.

U2R: unauthorized access to local root privileges.

probing: surveillance and another probing.

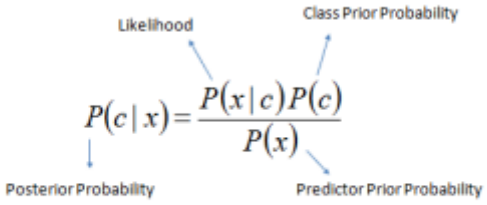
There are 21 types of attacks, in each category has various attacks the data is processed before it is used in a model. The features of the dataset have to be selected at first. The dataset does not provide the best performances from the intrusion detection system while including all the features. It results an increase in the computational cost and also the error rate of the system. This may due to the features are redundant or are not useful for making a division between different classes.

IV. RESULTS AND DISCUSSIONS

The two phases to build a Machine learning model is: training and experimenting. The machine learning process starts with a training phase. Training a dataset teaches the algorithm how to make accurate predictions. In the training phase, the machine learning model is developed. The model is trained and tested thus, it helps to validate the training method and improve the results.

4.1 Naive Bayes Classifier

Naive bayes algorithm is a supervised learning algorithm. It is the simplest and most effective classification algorithms. This can be implemented for both binary as well as multi-class classifications. Naïve bayes classifier can build fast machine learning models to make quick prediction. The occurrence of a particular feature is independent to the presence of other features is the assumption of this algorithm. The name Naive is assigned such that the algorithm assumes that the attributes are conditionally independent. This classification algorithm is based on the Bayes' theorem.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

From the above equation, P(c|x) is represented as the posterior probability of class (c, target) given predictor (x, attributes). P(c) is the prior probability of the class. P(x|c) is the likelihood that is the probability of the predictor given class, and P(x) is the prior probability of predictor.

It works based on the bayes theorem to verify the class of unknown dataset. The presence of a particular feature is unrelated to the presence of any other feature is naïve bayes. It is useful for very larger datasets and requires small amount of training data. Naive Bayes can struggle with zero probability problem, when the conditional probability is zero for a particular instance, then the model will be unable to make a prediction. One of the solutions is to use a smoothing procedure. A simplest smoothing technique is Laplace estimation. The three types of Naive Bayes model under the scikit-learn library are: (a) Gaussian naïve bayes, which assumes features follow a bell-like, normal distribution (b) Multinomial used for discrete counts, with respect to the quantity of times an outcome is observed across x trials and (c) Bernoulli useful for binary feature vectors; popular use-case is text classification.

4.2 Decision tree

Decision tree is a supervised machine learning algorithm and is used for both classification and regression problems. This algorithm applies the series of decisions with the given dataset. The model has a standard tree structure with nodes, branches, and leaf. An attribute or a feature is the nodes of decision tree. A decision or a rule is represented by the branch while each leaf represents a possible outcome or class label. The DT algorithm is able to select the best features for building a tree. The pruning operation is performed to remove the irrelevant branches from the tree to avoid the overfitting.

The decision tree algorithms such as CART, C4.5, and ID3 are most commonly used. Many ensemble learning algorithms like Random Forest (RF) and XGBoost are developed from different decision trees [26]. The Decision Trees are easy to understand and visualize. It can also handle both numerical data and categorical data. DT checks a condition, if it is true moves to the adjacent node with the decision. On the other hand, complex trees do not generalize well and decision trees are somewhat unstable resulting small

variations in the data might result in a completely different tree being generated.

The tree is constructed in a top-down recursive divide and conquer approach. A decision node will have more than two branches and a leaf represents a classification or decision. The topmost node in the decision tree that relates to the best predictor is called the root node, and the best thing about a decision tree is that it can handle both categorical data and numerical data.

4.3 Logistic Regression

Logistic Regression represents a classification model. It is used on binary classification problems with multiple classes, relating with the multinomial and ordinal logistic regression. Logistic Regression predicts categorical independent variable using a set of independent variables. Logistic regression is used to solve classification problems. It provides probabilities. Continuous and discrete datasets are used to classify new data. It is easy to identify the most effective variables used for classification.

The logistic regression finds a best-fitting relationship between the dependent variable and a group of independent variables. It is better than other binary classification algorithms like k nearest neighbor since it quantitatively explains the factors leading to classification. Here's what the logistic equation looks like:

$$Z_i = ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_1 x_1 + .. + \beta_n x_n$$

Taking e (exponent) on RHS and LHS of the equation results in:

$$P_i = E(y=1|x_i) = \frac{e^z}{1+e^z} = \frac{e^{\alpha+\beta_i x_i}}{1+e^{\alpha+\beta_i x_i}}$$

Logistic regression is specifically meant for classification, it is useful in understanding how a group of independent variables affect the outcome of the dependent variable.

4.4 Ensemble Model

The ensemble method combines multiple models to create one ensemble model [24]. The model combines the performances from multiple models to increase the overall performance and the accuracy could get boosted. The ensemble learning techniques are bagging, boosting and stacking. Ensembles are predictive models that combine more than one models and build a final ensemble model [25]. It often produces more better outputs than a single model can perform. The ensemble learning combines the classification problem or a regression problem.

In this work, different homogeneous models are selected as weak learners and are combined together. Each of the weak learners has its own output on the entire training set or on a fraction of the entire training set in the problem. The results of each weak learner are combined together to obtain the final output.

4.5 Hybrid Model

Hybrid Machine Learning forms a delectation of the machine learning algorithms that flawlessly combines different algorithms, processes, or procedures from similar or different fields of application with the objective of complementing each other. No single machine learning method is applicable to all problems. Some methods are good in handling noisy data but it may not be able to handle high-dimensional input data. Some others may be well on high-dimensional input data but may not be able to handle necessary data. Thus, the hybrid machine learning is implemented to complement the candidate methods and use one to overcome the weakness of the others.

Hybrid method combines two or more machine learning methods. These methods lead for higher performance and optimum results. Hybrid methods has the advantage of one or more methods reach better performance. To get accurate output the hybrid methods contain one unit for prediction and one unit for the optimization of the prediction unit.

The figure below is the representation of a hybrid model. The input data is passed through different classifiers. Each classifier is processed then predict the output and passed to the hybrid classifier. The hybrid classification can be done with respect to the classifier assigned the outcome of individual classifier is integrated and predicts the output

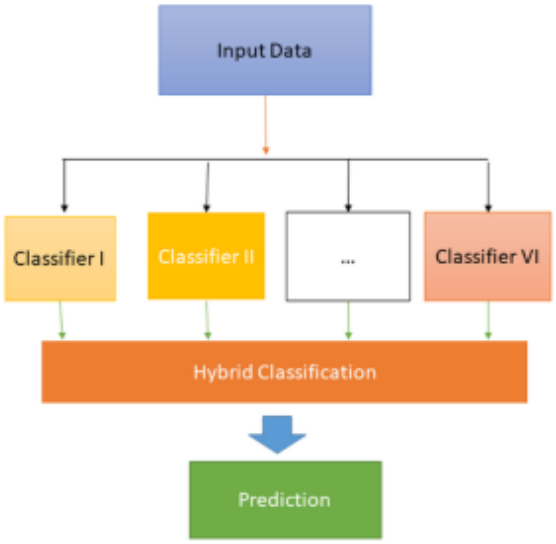


Fig. 2. Hybrid model representation Comparison results

The hybrid methods contain individual methods of different algorithms and form a method with higher flexibility and also high capability compared with single methods. Hybrid methods have high potential and capability than individual models thus it become more popular [17]. Table below shows different datasets and different classifiers selected by the investigators to find the accuracy value of the hybrid model

Table 1. Studies developed by hybrid model

Reference	Approach	Accuracy(%)	Dataset
Tang	DBN+LR	97.0	KDD 99
Qatf	SAE+SVM	93.96	KDD99
Qatf	Deep VAE	84.96	NSL KDD
Farahnakian	AE	94.71	KDD99
Naseer	DNN	89.0	NSL KDD
Bandyopadhyay	DCNN	84.58	UNSW NB15
Albahar	ENSEMBLE	93.3	ISCX-2012
Monshizadeh	MCA+EMD	87.29	ISCX-2012
Thi-Thu	FS+DT	95.33	ISCX-2012
Mighan	SAE+SVM	90.3	ISCX-2012
Wang	HAST-IDS	96.6	ISCX-2012

A hybrid model combines two or more machine learning model for higher performance and to get optimum results. The graph below shows the growth of hybrid and ensemble models.

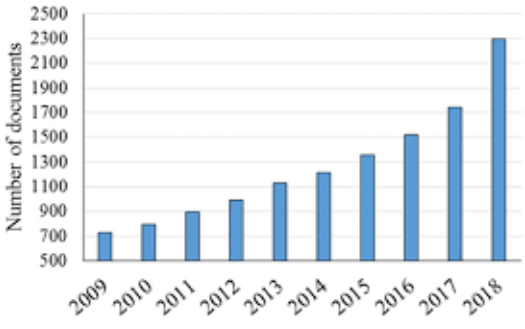


Fig. 3. The growing trend of hybrid and ensemble ML methods (source: web of science).

In our work, a diversified group of weak learners is combined to develop a hybrid ensemble learning model. The heterogeneous collection of models makes the hybrid model. The hybrid ensemble model of weak learning is applied to the kddcup99 dataset.

4.5 Confusion Matrix

A confusion matrix makes the prediction result of a classification method. It describes the production of a classification model. It calculates recall, precision, accuracy, true positive and false positive values. In a binary classifier the confusion matrix shows four different combinations of predicted and actual values in the case for a binary classifier. A true positive result that the model correctly predicts the positive class. A true negative predicts an output of the model which correctly predicts the negative class.

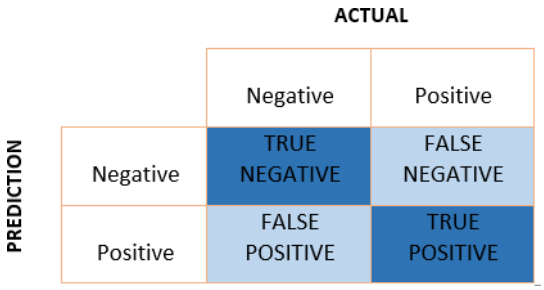


Fig. 4. Confusion matrix

The figure represents the confusion matrix of a binary classifier. The actual class with prediction class representing four outcomes is represented. The performance metrics is described below.

4.5 Performance Metrics

The performance metrics is the method for calculating the performance of machine learning algorithms for IDS. All the evaluation metrics are formulated on the different attributes used in the confusion matrix, which is a two-dimensional matrix with the data about the Actual and Predicted class and includes;

i. True Positive (TP): The data samples correctly predicted as an attack by the classifier.

ii. False Negative (FN): The data samples wrongly predicted as normal instances.

iii. False Positive (FP): The data samples wrongly classified as an attack.

iv. True Negative (TN): The samples correctly classified as normal instances.

The diagonal of confusion matrix represents the correct predictions of the classifier and the nondiagonal elements are the wrong predictions of a classifier. The evaluation metrics used basically are described below:
Precision: It is the ratio of correctly predicted attacks to the total samples predicted as attacks.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

Recall: It is a ratio of all samples correctly classified as attacks to the total samples that are actually attacks. This is also called as Detection Rate.

$$\text{Recall} = \text{Detection Rate} = \frac{TP}{TP + FN}.$$

Accuracy: It is the proportion of correctly classified instances to the whole number of instances. It is also called as Detection Accuracy and is a useful performance measure only when a dataset is balanced.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

F-Measure: It is determined as the harmonic mean of the Precision and Recall. It is a statistical technique for examining the accuracy of a system by considering both precision and recall of the system [21].

$$\text{F Measure} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right).$$

4.5 Voting Classifier

A Voting Classifier is a machine learning model that trains on ensemble models and predicts an output with respect to the highest probability of selected class as the result. It manages the independent variables of the training dataset with the dependent variables. After fitting, it makes predictions and the accuracy of its predictions is evaluated [13].

The two easiest ensemble methods are voting and averaging. Both are easy to evaluate and implement. Voting is used for classification problems and averaging is used for regression methods. The primary step is to develop multiple classification or regression models using the training dataset. Each base model can be developed with multiple modules of the same training dataset and same algorithm, or using the same dataset with different algorithms, or any other method.

4.6 Variance

Variance is the measure of a model's sensitivity to fluctuations in the data. A model may learn from noise. It leads a model to consider not important features as important. If the variance is high, the model will contain all the attributes of the given data to it, will maintain itself to the data, and predict on it well but new data may not have the same features and the model is not able to predict on it well. This is called Overfitting [3].

4.6 K-Fold Cross Validation

Over-fitting is the general problem prevalent in most of the machine learning models. K-fold cross- validation can be evaluated to check if the model is over-fitted at all. In this validation, the data set is arbitrarily partitioned into k mutually exclusive subsets, each of which is of the same size. Out of these, one is retained for testing and others are used to train the model. The same process is done for all k folds.

Cross-validation is to establish that the given model does not overfit the training set and it also improves the versatility of the model. Thus, the model can be used to predict the labels in the testing dataset. The predicted labels are then differentiated to the actual testing set labels through metrics such as confusion matrix, precision score, recall score, F1-score, roc auc score.

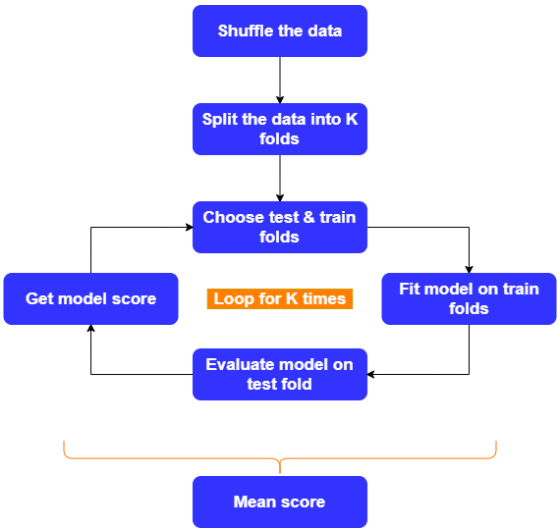


Fig. 5. Implementing cross validation

In k fold cross validation, data set is divided into k subsets and the process is repeated k times. Each time a subset is used one act as test set and others as train set. Then the average accuracy across all k fold validation is computed.

4.7 Roc-Auc score

Receiver operating characteristics is a visual comparison of classification models, shows the connection between the true positive rate and the false positive rate. ROC curve is an important classification performance evaluation metric. The ROC curve shows the relation of the classifier by organizing the true positive rate to the rate of false positives. For an outstanding classifier, the true positive rate will increase [19].

Table 2. Roc-auc scores of individual models

ALGORITHM	ROC-AUC SCORE	ACCURACY
LR model	0.861	0.943
DT model	0.975	0.987
Naïve bayes	0.932	0.965

Table1 shows the roc-auc score with accuracy value of three machine learning models. The decision tree having higher accuracy and higher roc-auc score. The logistic

regression model has less accuracy value and also the roc-auc score is less than other two models. The roc curve is represented by plotting the false positive rate in y-axis with respect to true positive rate in x-axis.

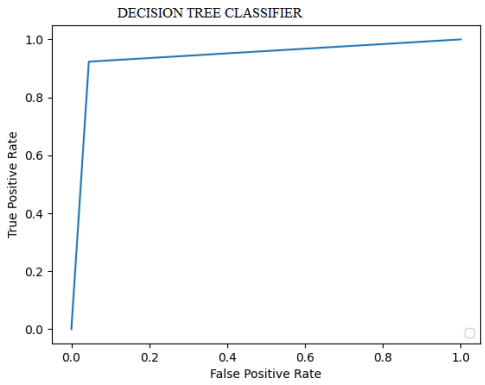


Fig. 6. Roc curve of decision tree classifier

The Roc curve of the decision tree model is shown above. The false positive rate is in x-axis and true positive rate is marked in y-axis. In python, roc curve can be drawn by importing the metrics library with roc modules from sklearn. The roc-auc score is0.975 for the decision tree classifier.

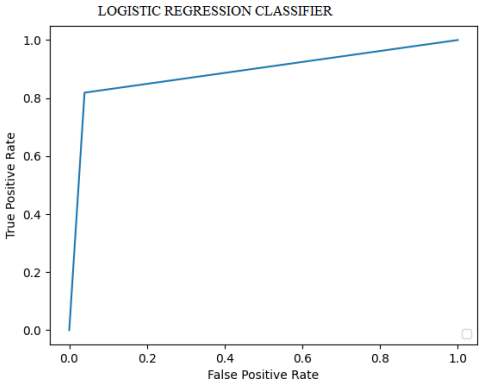


Fig. 7. Roc curve of logistic regression

The logistic regression has a low roc-auc score as 0.861. Comparing with other two machine learning models the accuracy value also low. The LR model is very efficient to train. It is implemented for binary and linear classification problems. The roc -auc score of naive bayes is represented below. The value is 0.932. The naïve bayes model is simple and easy to implement. It is fast and can be used for real-time predictions.

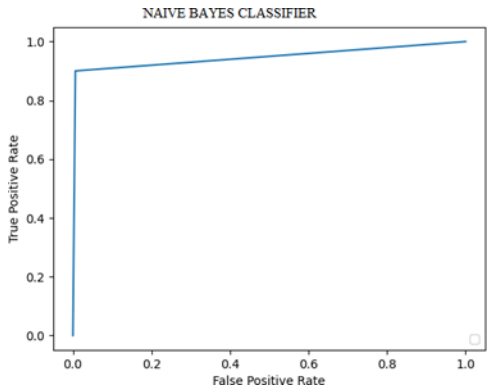


Fig. 8. Roc curve of naïve bayes classifier

The graphical representation shows the accuracy values of individual classifiers. Accuracy is the simple ratio between the correctly classified instances with the total number of instances.

Precision is the ratio of number of positive classes which is actually belongs to positive class. It is the ratio of true positives with all positives. Recall is theratio of number of positive class prediction with total instances. The model correctly identifies true positives is recall. Here is the tabular representationof true positive with precision for the threeclassifiers.

Table 3. True postive v/s precision of individual model

Algorithm	True positive rate	Precision
Naïve bayes	0.987	0.970
Decision Tree	0.994	0.991
LR Model	0.996	0.937

The hybrid ensemble model here integrates the logistic regression and decision tree. The two classification methods are combined together and evaluated the performance metrices. The accuracy, precision, recall and f1score is evaluated. The accuracy of this hybrid model is evaluated as 0.989. The hybrid method improves the accuracy of the model. The efficiency of a classifier can beimproved through these ensemble methods. The model is trained and tested first. The k fold cross validation is done to the model to test the ability of the model to predict new data. The ensemble method integrates the base model to predict one optimal model. Thus, evaluating the performance metrics of the model.

Table 4. Evaluating performance metrics of LR and DT

Performance Metrics	Hybrid Ensemble Model (LR+DT)	Variance value
Accuracy	0.989	0.12
Precision	0.999	0.00
Recall	0.978	0.03
F1 Score	0.988	0.01

In this work, the classification accuracies are evaluated using the confusion matrices of individual models. The 10-fold cross-validations are performed and evaluated for more validation of the accuracies of proposed models. Each model is defined 5 times thus a combination of a total of 15 weak learners is formed. Then finally, the Max Voting Classifier method is evaluated. The hybrid ensemble learning model has outperformed other than the individual learning model. We can observe this by verifying the performance metrices of ensemble model with respect to individual model. In future, we can develop more hybrid ensemble models by combining the weak learners.

The accuracy score of two ensemble methods is represented. This research work helps to conclude that the ensemble method can make better performance for the classifier, it can tolerate data incompleteness.

Table 5. Evaluating Performance metrics of LR+DT+NB

Performance Metrics	Hybrid Ensemble Model (LR+DT+NB)	Variance value
Accuracy	0.985	0.16
Precision	0.989	0.01
Recall	0.977	0.03
F1 Score	0.983	0.02

The table above shows the experimental results done by the hybrid ensemble method, integrating three major machine learning models. The variance value evaluates the output of the classifier. Less variance value means better performance model. From the above table we can see that the variance value is less thus, our model is well performed.

V. CONCLUSION

Machine learning uses algorithms to learn from the data, and make intellectual predictions according towhat it has learned. We have trained three machine learning models including, logistic regression, decision tree and naïve bayes. The performance metrics of hybrid ensemble model is evaluated. Themodel with decision tree, logistic regression and naïve bayes results 0.985 and another model combining logistic regression with decision tree obtained an accuracy of 0.989. Analyzing the results, we can reach the conclusion that the hybrid ensemble model makes better accuracy rate than individual classifiers. In the future, the hybrid modelof unsupervised learning can be done.

References

[1] Mohammad Almseidin, Maen Alzubi, Szilveszter Kovacs and Mouhammd Alkasassbeh, "Evaluation of Machine Learning Algorithms for Intrusion Detection System", 1801.02330.pdf (arxiv.org)

[2] Why Use Ensemble Learning? (machinelearningmastery.com)

[3] Zina Chkirbene, Sohaila Eltanbouly, "Hybrid Machine Learning for Network Anomaly Intrusion Detection",2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT).

[4] Muhammad Ashfaq Khan and Yangwoo Kim, "Deep Learning-Based Hybrid Intelligent Intrusion Detection System", Computers, Materials &Continua DOI:10.32604/cmc.2021.015647.

[5] Harsh H. Patel, Purvi Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", International Journal of Computer Sciences and Engineering.

[6] Hatim Mohamad Tahir, Wael Hasan, "HYBRID MACHINE LEARNING TECHNIQUE FOR INTRUSION DETECTION SYSTEM", Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015 11-13 August, 2015, Istanbul, Turkey S. Bandyopadhyay and et al., A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data, IEEE/ACM TCBB, 2014, DOI: 10.1109/TCBB.2013.147.

[7] Hind Bangui, Mouzhi Ge, "A hybrid machine learning model for intrusion detection in VANET", SpringerLink.

[8] Po-Jen Chuang and Si-Han Li, "Network Intrusion Detection using Hybrid Machine Learning", 2019 International Conference on Fuzzy Theory and Its Applications (iFUZZY)

[9] Tang, TA, Mhamdi, L, McLernon, "Deep Learning Approach for Network Intrusion Detection in Software Defined Networking", IEEE Conference Publication | IEEE Xplore

[10] Ünal Çavuşoğlu, "A new hybrid approach for intrusion detection using machine learning methods",SpringerLink.

[11] Abebe Tesfahun, D. Lalitha Bhaskari, "Effective Hybrid Intrusion Detection System: A LayeredApproach", I. J. Computer Network and Information Security, 2015, 3, 35-41

[12] How VOTing classifiers work! A scikit-learn feature for enhancing… | by Mubarak Ganiyu | Towards DataScience.

[13] Rahul Vigneswaran K, Vinayakumar R, "Evaluating Shallow and Deep Neural Networks for Network Intrusion Detection Systems in Cyber Security".

[14] S. Angra and S. Ahuja, "Machine learning and its applications: A review", 2017.

[15] Samriddhi Verma, Nithyanandam P, "Detailed Analysis of Intrusion Detection using Machine Learning Algorithms", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277- 3878, Volume-9 Issue-1, May 2020.

[16] Sina Ardabili, Amir Mosavi, "Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods", doi: 10.20944/preprints 201908.0203.v

[17] Suad Mohammed Othman, Fadl Mutaher Ba-Alwi, "Intrusion detection model using machine learning algorithm on Big Data environment", Othman et al. J Big Data (2018) 5:34 https://doi.org/10.1186/s40537-018-0145-4

[18] Supervised Machine Learning Classification: An In-Depth Guide | Built In Machine Learning & Training Data: Sources, Methods, Things to Keep in Mind (labelyourdata.com)

[19] Module 21 - How to build a Machine Learning Intrusion Detection system - Blue Teams Academy - Free Training 2021.

[20] Valeriy Gavrishchaka, Zhenyi Yang, "Advantages of Hybrid Deep Learning Frameworks in Applications with Limited Data", International Journal of MachineLearning and Computing, Vol. 8, No. 6, December 2018.

[21] Varsha P.Desai, Dr.K.S.Oza, "Data Mining Approach for Cyber Security", International Journal ofComputer Applications Technology and ResearchVolume 10–Issue 01, 35-40, 2021, ISSN:-2319–8656

[22] Ensemble Methods in Machine Learning: What are They and Why Use Them? | by Evan Lutins | TowardsData Science

[23] Why UseEnsembleLearning? (machinelearningmastery.com)Zeeshan Ahmad, Adnan Shahid Khan, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches", https://doi.org/10.1002/ett.4150.

[12] What Is the Difference Between Data Mining and Machine Learning? | Bernard Marr

[24] Ahiya Ahammed, Balazs Harangi, Andras Hajdu, "Hybrid AdaBoost and Naïve Bayes Classifier for Supervised Learning", published at http://ceur- ws.org, November 6–8, 2020

[25] Classification in Machine Learning | The BestClassification Models (simplilearn.com)