ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse99942020.pdf

https://doi.org/10.30534/ijatcse/2020/99942020

Hybrid Framework for Intrusion Detection System using Ensemble Approach

***S. R. Khonde[1], V. Ulagamuthalvi[2]**

[1]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
*khonde.shraddha@gmail.com

[2]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
ulagamv@gmail.com

ABSTRACT

Malicious attack detection is a new emerging area of research now days due to huge number of internet and network usage. Attack detection in network is handled by a system called Intrusion Detection System (IDS). Most of the administrators make use of IDS for monitoring malicious activities of the network. To increase attack detection rate for securing network IDS need to work intelligently. Various machine learning algorithms are used to improve IDS performance considering threat of attacks in modern era of internet. In this paper a novel framework is proposed which will make use of signature as well as anomaly based detection to increase detection rate and reduce false alarm rate. This architecture makes use of various supervised and unsupervised machine learning algorithms for testing real time internet traffic. Dataset used for testing proposed framework is Intrusion Detection Evaluation Dataset CICIDS-17. This framework emphasis of attack detection using signature based detection and propose a new method for new attack detection using anomaly based identification. Dataset used for training deals with various modern attacks and helps to find signature of new attack with help of 88 features of dataset. Various feature selection techniques are used to reduce number of features from dataset to reduce computation time of the system. As this framework is proposed for distributed networks feature selection plays a vital role in performance of system. An experiment results shows that proposed architecture which makes use of ensemble approach provides better performance in terms of detection rate and false alarm rate. Proposed architecture shows increase in detection rate by 5% for signature based detection and 2% for anomaly based detection. Reduction of 0.05 is observed in false alarm rate.

Key words: Ensemble classifier, Intrusion Detection System, Random forest, Isolation Forest, SVM, ANN.

1. INTRODUCTION

Modern era provides ease to various users for making use of huge facilities provided by internet. While using internet facilities users share huge amount of data across internet. Any form of data can be used and shared on internet while using it. Most of the organizations including individual users are making their private data available on internet. This can create a big threat for network. Most the malicious activities can increase threat to network security. Due to increase in data transfer through internet, network security in a wide area of research. Most of the networks are facing problem in maintaining privacy, confidentiality and security of data. To increase security of such network most of the network administrators and organizations use security tools. Tools like firewall, user authentication, access control, anti-virus etc. are used to avoid malicious activities to happen in network. These tools provides security to network but not able to secure network from internal malicious activities. Most of these tools are not as efficient in handling intrusions entering into the system. However, IDS is an intelligent system which can monitor all activities happening inside and outside of the network. It monitors are data entering and passing through the network to avoid invalid activity. IDS informs administrator if any malicious activity is observed in the network such as huge amount of packets for single system, shutdown of systems in network, trouble shooting, system not responding to activities etc. Administrator can take appropriate action on these activities to secure network and avoid data loss in network. IDS usually helps administrator to observe, monitor and rectify malicious activities in network. Intruders use to find loop holes in various security tools used to provide security to network and can easily get access of system inside the network. After getting access to network system they can harm confidential information related to networks and can change or alter confidential data. To avoid loss of data and entry of intruder in the network an intelligent device is required which can take decisions on his own for avoiding losses in network. Such system can be implemented using various data mining, machine learning and neural network algorithms. It should work in an intelligent way such that old as well as new attacks can be detected before entering in the system.

According to research, attacks also called as malicious activities can be detected using signatures or anomaly based detection. Signature based detection is based on signatures of attacks stored in the network. Whenever any data in form of packet enters in the system, this type of IDS matches all attack signatures to it. It matches found then data is considered as malicious activity and rejected from the network. If match does not found then data is allowed to enter in the network. Basic limitation of this method is signature. If signature of a particular attack is not available in dataset then IDS will not able to detect that attack and network data can be harmed. To overcome this limitation dataset consist of signatures of modern attacks need to consider for training and testing IDS while implementation. Once the packet enters in the system it is checked by anomaly based detection system if any abnormal behavior is observed in the network. Anomaly based detection is based on behavior of the system. If IDS observe change in normal behavior of the network or system it considers it as anomaly or new type of attack and rejected from the system. As the packet flows in the network anomaly based IDS continuously observe state of network to monitor any abnormal activity happening in the system. Such packet and its behavior are captured and stored in form of signature in the dataset so that in future packet with same behavior can be caught while signature based detection.

Network and user system are secured by IDS from getting compromised from intruders. It also maintained various security policies such as Confidentiality, Availability and Integrity. Violation of security policies can be avoided by IDS. In modern era attacks are emerging in the network with new form. Some of those are old means well known and some are new that is unknown. Abnormal behavior of a system identifies attack. Attack detection in IDS is done using two ways as Signature based IDS where signatures from known data-set is used to match signatures of various attacks. Another is Anomaly based IDS which observe and monitor behavior of system. Both detection methods have own strength and weakness. In known attack environment Signature based IDS provide better accuracy but not for unknown attacks. Whereas Anomaly based detection accuracy is less as to find normal behaviors of system is difficult and most of the time normal packets are also considered as attack. Little deviation in the normal behavior can be considered as new attack. This leads toward increase in false alarm rate (FAR).

From various researches in this area it is observed that both detection methods are useful in attack detection provided used in proper fashion. To makeup limitation of both methods in proposed framework both methods are used in two phases to avoid any intruder entering inside the system. These both are used in hybrid distributed manner to monitor the state of network and monitor all activities happening. This framework avoids limitation of both methods to some extent. Signature based detection method is used to identify all the known attacks and Anomaly based detection is used for detecting new attacks by proposed method. Intrusion Detection Evaluation Data-set CICIDS-17 is used for training and testing proposed framework. Many classifiers are used for training and testing such as random forest, neural network, artificial neural network and isolation forest. All these classifiers performance is compared with ensemble approach of all classifier to prove the efficiency of proposed framework. This data-set consist of total 88 features and can be trained for most of the known attacks such as probe, phishing, zero day attack, DoS, DDoS, U2R, R2L etc. A feature selection technique is used for dimensionality reduction to improve computation time of system. Evaluation of proposed system is done using various performance parameters such as detection rate, accuracy, precision and false alarm rate. Experimental results prove that proposed hybrid framework increases detection accuracy, detection rate, precision and reduces false alarm rate.

2. LITERATURE SURVEY

IDS available now days for securing network make use of single classifier for detecting attacks entering in the network. As every classifier has its limitation sometimes prediction of attack gives biased output and not able to detect some on malicious activities. Mostly researchers suggest using ensemble classifiers to find abnormal activities in the network to reduce number of false alarms and increase accuracy. It is proved by many researchers that using classifiers in ensembling approach provides better performance of IDS as compared to individual classifier.
Multi-Classifier systems have received much attention due to their focus on combining the output of classifiers and create a final decision. To achieve best possible results, the proposed hybrid IDS use combination of signature-based systems and anomaly-based systems.

A hybrid approach which uses 2 stages is elaborated in [1]. In this misuse and anomaly attacks were detected. K means neighbor algorithm is used for signature based detection and k means clustering for anomaly detection. Using this approach detection accuracy obtained is of 95.76% and false alarm rate reduced by 1.05%. Most of the researchers work to reduce false alarm rate using techniques such as feature reduction and dimensionality reduction. A novel method for reducing FAR is elaborated in [2]. Evaluation parameters used are high alert frequency, neighboring related alert and normal false positives. These components are used to find the source IP and time interval in which source is sending same type of malicious packets. Arrival intervals are also observed to identify source of the attacker. As the experimental results proposed method increase accuracy and reduce FAR by 90%. Ensemble approach importance is elaborated in [3]. Various ensemble approaches which can be used for improving performance of IDS is elaborated in detail. Performance of IDS improves drastically as compare to single classifier detection.
An anomaly based approach is proposed which does not use any dataset to train their algorithm [4]. Steps are used for construction as Multi Resolution Flow Aggregation method Sub Space Clustering method and simple threshold detection approach. Attack is detected based on outlier threshold, if value goes beyond specified, it is declared as attack. This proposed method has reduced FAR and increased accuracy.
Ensemble approach by combining various supervised algorithms is explained in [5]. This paper proposed approach to deal with unbalanced data from dataset while detection. Model use ensemble of k means, support vector machine and decision tree. Final prediction is taken from

combination of vote from all classifiers. Results show detection accuracy of 99.2% for normal behavior, 98.21% or denial of service, 93.22% for probe, 44.44% for U2R and 93.21% for R2L attack detection. Ensemble classifiers SVM, RF and NN are used for attack detection in [6]. Comparison provided in paper proves that ensemble approach provides better detection accuracy and reduces FAR as compared to single classifier. Author makes use of feature selection techniques to improve performance of IDS. New approaches using emerging concepts like artificial intelligence, neural network, deep learning are used by many researches towards increasing accuracy and efficiency of IDS. A new approach using benchmark dataset NSL-KDD is elaborated in [7]. Feature selection is used to select best attribute from 41 features of this dataset. This paper mostly shows the importance of feature selection for improvement in performance of system. While testing this dataset is divided into four parts as basic, traffic, host and content. Training and testing is done using random forest classifier. For all parts tress of RF is created to test accuracy. Results show that feature selection plays a vital role in improving performance of IDS. A hybrid model using SVM and DT in ensemble is explained in [8]. This paper focused on ensemble techniques used for attack detection. It also explained importance of ensembling classifiers rather than using single classifier for detection. Each classifier provides wrong output in some cases, so we cannot depend on single classifier for final prediction. This issue of single classifier can be easily solved using ensemble classifier. Advantages of ensemble classifier long with survey of various ensembling techniques used in homogenous and heterogeneous network is explained in [9]. A model for analyzing normal and malicious traffic is explained in [10]. Detection is done with the help of deep packet inspection. Paper explained model with two hidden layers with 30 nodes each with 1000 input and 2 output nodes. Model provides accuracy of 98% with 97% precision and 95% sensitivity. A model based on feed forward neural network is explained in [11]. Results prove that as detection rate is increased and false alarm rate is decreased when feed forward network is used. Dataset used is KDD which is split into 10% training and 90% testing. 84.12% accuracy is obtained by model without dataset split and 64.42% after splitting dataset. All unlabeled instances are labeled using divide and conquer strategy in this model. A pseudo bayes estimator model is explained in [12]. It is used to find probabilities of new attacks penetrated in the network in real time environment. This probability of new attack is analyzed by naïve bayes classifier. This classification is used to classify network traffic into known and new attacks.

A survey of various IDS system classifiers which used KDD 99 dataset for experiments is explained in [13]. Supervised and unsupervised algorithms are explained in this paper. Results show that supervised algorithms such as DT and SVM provides better results than unsupervised algorithm. Decision tree provides more accuracy than support vector machine. A novel modular ensemble method is proposed in [14]. Method focused on web related services for attack detection such as mail service, web service etc. Each service is handled by single classifier whose output is then ensemble together for finding final prediction. Ensemble is done with the help of rule like maximum, minimum, mean and product rule. KDD 99 dataset is used for training and testing for all classifiers. Ensemble techniques with various classifiers are elaborated in [15]. SVM and KNN are used as base classifiers in this model. Weighted majority voting algorithm is used for ensembling various classifiers. To use this algorithm weights need to be assign by classifier. Weight assignment is done with help of three techniques as particle swarm optimization (PSO), variant of PSO which uses iterative sampling and weighted majority algorithm. Functionality of weighted majority algorithm is elaborated in [16]. To improve performance of this algorithm feature selection technique can be used. Authors used PSO variant for giving better results with weighted majority algorithm. An architecture based on neural network is explained by authors in [17]. Ada-boost ensembling method is used for combining output from base classifiers. Some samples are created and neural network classifiers are trained to find fitness function, which can be used to validate set of samples trained. Experiment results proves that ensemble detection has more efficiency in terms of recall, precision, f-measure, true positive rate an false positive rate as compared to single classifier.

3. METHODOLOGY

The proposed framework is proposed for networks working in distributed fashion. In system where nodes are connected with each other each node will consist of ensemble of classifier used for detection of attack at each node of network. Proposed framework makes used of signature as well as anomaly based detection in the network for detection of attacks. Architecture is testing on real time data captures from the network. Classifiers used are random forest, isolation forest, neural network and artificial neural network. Using this framework network data will be scanned twice once while signature based detection and other with anomaly based detection. All the classifiers used are trained using dataset CIDIDS-2017. This dataset consist of signatures of various modern attacks. Most of the standard datasets are available for attack detection which can be used for IDS. Reason behind considering this dataset for implementation of this dataset is this dataset is created for modern attacks using real time traffic capture. Each data packet entering in the system will be analyzed first with signature based detection. This analysis is done with the help of signatures stored in the dataset. Packets data is matched with various categories of attacks signatures. If signature is matched then data will not be allowed to enter in the network and rejected by the node. If matching signatures are not available and matching patterns are not retrieving from dataset then packet is accepted and passed inside the network. As packet passes form one node to other packet behavior is monitored by each node to check any abnormal activity in the network. This is the second phase of architecture which works on anomaly based attack detection. In this method all classifiers are trained to maintain normal behavior of systems. Any deviation from this behavior is considered as malicious activity. If any change in normal behavior is observed node prepares a log of packet behavior and stored it for further processing. Packet is then immediately rejected by network to avoid harm of data in network. In next phase this log is used to create the signature of new unknown attack so that it can be used in first phase that is in signature based detection before entering inside the network. Updating of the dataset is informed to all nodes to update signatures. All nodes are using ensembling of RF, IF, SVM and ANN classifiers. Ensembling is done using weighted majority algorithm. Ensembling is used for improving detection accuracy of classifier in signature as well as anomaly based detection. All classifier are trained for signature as well as anomaly based detection so that framework can be tested in real time environment. Hybrid framework is represented in Figure 1.

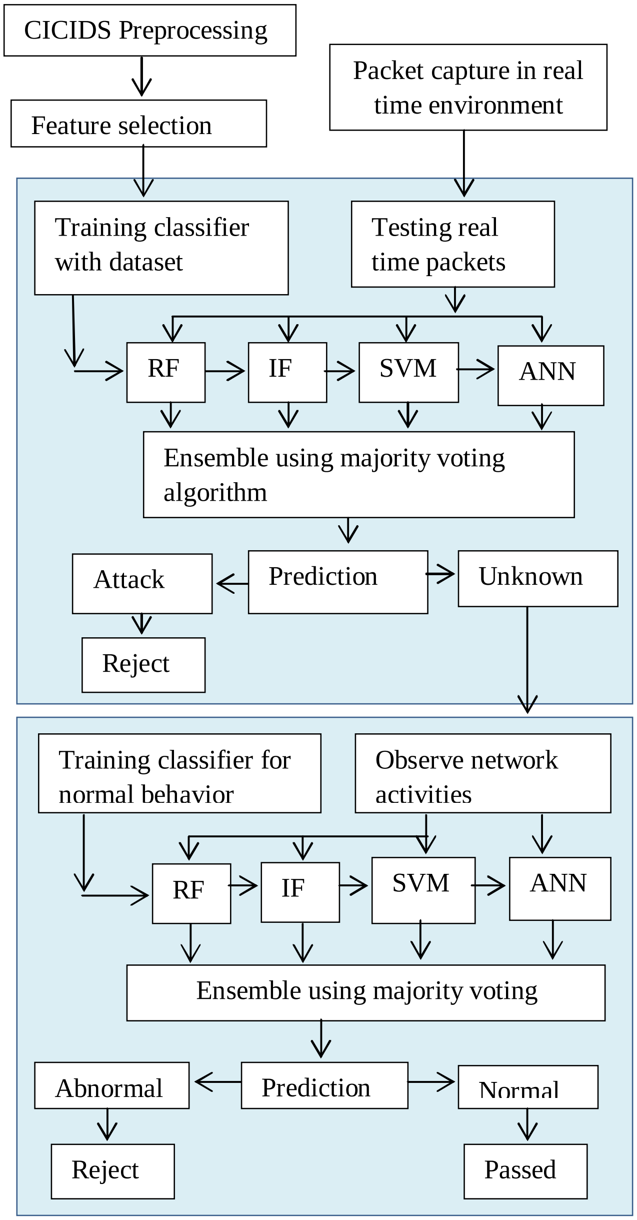
CICIDS Preprocessing
Feature selection
Packet capture in real time environment
Training classifier with dataset
Testing real time packets
RF
IF
SVM
ANN
Ensemble using majority voting algorithm
Attack
Prediction
Unknown
Reject
Training classifier for normal behavior
Observe network activities
RF
IF
SVM
ANN
Ensemble using majority voting
Abnormal
Prediction
Normal
Reject
Passed

Figure 1: Architecture of Hybrid Distributed IDS

Dataset used for training and testing is ISCX CCICIDS 2017 which consist of various modern day attacks. Detailed explanation of dataset, feature selection techniques used and ensembling of classifiers is explained in followed sections.

3.1 CICIDS- 2017 Dataset

Many standard datasets are available for intrusion detection system. Some of them are KDD 99, DARPA, UNSW, CICIDS and many more. All these datasets are having signatures of well know attacks. Some are them like KDD99 and DARPA are old datasets which does not contain signatures of new modern attacks. This is the main reason behind using CICIDS dataset for training and testing proposed system, as it contains signatures of modern and up to date attacks. To check performance of IDS in real time environment it should contains signature of various modern day attacks, which are not available in old datasets. CICIDS dataset is created using real time traffic that is the reason it stores signature of many modern attacks. Canadian Institute of cyber security made this dataset available for research in year 2017. This dataset consist of 83 features with total 3119345 instances of various attacks [18]. Institute captures real time data for one week to capture various new types of attacks on the network before generating this dataset. This data-set has in total 15 classes (1 normal label + 14 attack label). CICIDS 2017 is a data-set consist of many new real world attacks including DoS, DDoS, Brute Force, XSS, SQL Injection, Infiltration, Port scan, Botnet etc. Table 1 represents all types of attacks present in this

Table 1: Attack Class and Instance for CICDS2017 data-set

Class	Number of instances
BENIGN	2359087
DoS Hulk	231072
DoS GoldenEye	10293
SSH-Patator	5897
DoS Slowhttptest	5499
Bot	1966
Web Attack – Brute Force	1507
FTP-Patator	7938
DDoS	41835
PortScan	158930
DoS slowloris	5796
Heartbleed	11
Web Attack – Sql Injection	21
Web Attack – XSS	652
Infiltration	36

In proposed framework this dataset is used for signature as well as anomaly based detection in real time environment. To clean data stored in dataset and to normalize it data preprocessing is done. Important features are selected using feature selection technique to reduce overload on classifier while training and reducing computation time required for execution. All classifiers will be trained using reduces features of CICIDS dataset and tested in real time environment. Testing is done to find comparison of individual classifier against ensemble classifier.

3.2 Dataset Pre-processing

Preprocessing of dataset is required to clean and normalize dataset. This is considered as an important step before training classifiers using dataset. Dataset cleaning and removing of duplicate instances is done in this step so that training and processing time will be reduced during evaluation of performance for IDS. Dataset consist of some missing value instances this can reduce performance of attack detection. In such case use of this invalid values prediction of classifier may go wrong. To avoid it Min-Max normalization technique is used to normalize dataset. Data-set also consist of numeric values with huge range gap which need to be converted into some specific range. This is important to reduce the training time of the classifier. If wide range of values is used to train classifier it requires double of the time required with reduced range of values. Cleaning of the data-set is done for removing Nan, infinity and duplicate values and duplicate columns. Min-max normalization technique is used to normalize data-set numeric values. Equation 1 shows formula used for normalization of dataset.

$$x' = \frac{x - \min_A}{\max_A - \min_A}\left(n\_\max_A - n\_\min_A\right) + n\_\min_A \qquad (1)$$

Where
x' - is the normalized value of respective feature.
x – is the actual value of the features
$\min_A$ and $\max_A$ - is the actual minimum and maximum value of the feature
n_max_A and n_min_A - New normalized minimum and maximum value set for the feature.
A – Feature from the data-set

Features in dataset are represented using wide range of numeric values. To improve performance of IDS minimum and maximum values are set to -25 and +25 respectively for each feature. Equation 1 represents min-max normalization formula which is used to convert wide range of numeric data in the form of range i.e. -25 to +25 such that the summation will be 1 for each feature. To reduce training time required for classifier range for normalization is decided.

3.3 Feature Selection

Feature selection plays an important role while evaluating performance of the IDS system. As in proposed system number of classifiers is used to improve IDS performance and all nodes are connected in distributed fashion, it is mandatory to reduce number of features so that computation time and processing time can be reduced. By selecting only important features instead of all it helps to reduce training time, improve accuracy and reduces over fitting of data. As the data-set used in proposed architecture has 83 features it require to consider only important features which helps for attack detection and reducing training time of classifier. As proposed system works with multiple classifier time require to train classifier becomes the important factor in performance evaluation of IDS. To reduce over fitting of classifier feature selection techniques are used. Various feature selection techniques are available to reduce features of data-set. Some of the techniques are filter based and wrapper based such as chi-square [19], correlation, variable importance, particle swarm optimization, Euclidean distance [20] and many more. Proposed system makes use of correlation analysis for feature selection. This technique provides relation between all features available in dataset with each other. It is basically used for predicting behavior and relation between numbers of entities. In proposed system relation between features is used to find attack pattern form the data. For all features according to pattern features are selected to find correlation coefficient of each feature or attribute. The feature having coefficient value greater than equal to 1 and less than equal to 96 are considered as important features. These features are further considered for attack detection and training classifiers. Training of classifier is done with less number of features but testing is done in real time environment considering all features of the data-set. Person correlation coefficient is the method used to find coefficients of all features. It helps to measure and find the linear association between numbers of features of the data-set. Equation 2 shows formula for calculation correlation coefficient of each feature of the dataset.

$$r = \frac{\sum_i \left(p_i - \overline{p}\right)\left(q_i - \overline{q}\right)}{\sqrt{\sum_i \left(p_i - \overline{p}\right)^2}\sqrt{\sum_i \left(q_i - \overline{q}\right)^2}} \qquad (2)$$

Where p and q are the values for the features for which we need to find correlation.

Range for values of coefficient lies in between -1 to +1. Strong correlation between features is represented using +1 value or positive values. However, poor or negative correlation is represented using -1 and negative values. While selecting final features only positive correlation valued features are selected.
As per results presented in [21] it can be analyzed that for attack detection it is not mandatory to use all features of any data-set. Techniques like information gain and gain ratio gives less efficiency as compared to correlation coefficient analysis. In experiments authors used only 22 features instead of all methods for achieving same accuracy. Feature selection importance is elaborated in [22]. Various techniques which can be used for dimensionality reduction for classifiers are illustrated by author. The number of features reduced can be equivalent to √A if A is the number

of features present in the system. The scope for feature reduction is explained with given formula in [23]. After reduction less number of features can be used for attack detection in IDS and still it provides better accuracy. It also helps in reducing training and processing time of the classifiers used for detection. Any classifiers provide better results coming from machine learning or neural network when used with feature selection [24]. Feature selection importance [25] and various techniques like wrapper methods [26] are used.

3.4 Ensemble Classifiers

Proposed system makes use of ensembling technique to find final prediction about the data entering in the network. Final prediction can be normal or attack. Importance of ensembling in various networks is explained by [27]. Ensemble classification technique provides better accuracy when the IDS is build using feature selection technique as compare to using fill dataset [28]. Hybrid approach works in efficient manner and proved in [29]. In the proposed system method used for prediction of packet for attack detection is based on ensemble technique. Ensemble is basically used to combine results from number of classifiers for predicting final output. Proposed framework makes use of majority voting algorithm for combining output of various classifiers. The accuracy of detection increases when we make use of ensemble technique rather than single classifier. Classifiers can provide biased prediction sometimes can face over fitting problem in real time environment. Most of the researchers explain importance of ensemble on single classifier prediction. In proposed architecture ensemble is done with four classifiers as random forest, isolation forest, support vector machine and artificial neural network. These all classifiers are trained using dataset after reducing features from data-set.

Ensembling methods used are bagging, boosting and stacking. Homogeneous weak classifiers are combined using Bagging and Boosting. To increase capability of weak classifiers bagging ensembling is used. As proposed system is working in real time environment some classifiers can give biased prediction. To avoid it bagging ensembling can be used to increase capacity of each classifier. To boost classifier for improving its performance boosting ensembling method is used. The subset used in boosting is basically used to boost the accuracy and efficiency of the weak classifier. This method also provides better efficiency as compare to bagging. Heterogeneous weak classifiers are ensemble using staking method. Stacking method is used in proposed system as supervised and unsupervised classifiers are used in ensembling. Classifiers used in ensemble are explained in following subsections.

3.4.1 Random Forest Classifier

Random Forest is a supervised machine learning algorithm. Base idea for this classifier is taken form decision tree which works finely when used with labelled and static data. Limitation of DT is removed in RF as it makes use of multiple DT to form a forest of trees where each tree represents single DT. RF circumvents most of the limitations of DT as used in multiple numbers. Performance of RF is depending on number of trees used for training and testing classifier. The more number of trees formed in RF while training classifier more accuracy can be obtained. Effectiveness of RF in performance of IDS evaluation in represented in [30, 31]. To make RF robust for real time environment more number of trees is used [32].

3.4.2 Isolation Forest Classifier

To find abnormal behaviour of the classifier Isolation forest is used. It will train classifier in terms of normal behaviour such as any deviation in this performance will be used to generate attack on basis of abnormal behaviour. This is a supervised type of algorithm as before training classifier user should know normal behaviour features of network. This classifier can work in both supervised and unsupervised fashion depend on data availability. This classifier plays important role when it is used in anomaly based detection. Isolation forest mostly used to find the feature which mostly responsible for finding the anomaly values for the data. These trees are used to isolate the anomaly from huge dataset. To observe changes in huge dataset are difficult sometimes in such cases isolation forest helps us to find the outliers in data that is anomaly activities that are not normal. Isolation forest classifier mostly focused on finding anomaly rather than normal packets with help of tress build using this algorithm.

3.4.3 Support Vector Machine Classifier

Support vector machine is a supervised type of machine learning algorithm used for attack detection in IDS system. SVM works with binary data and provides amazing detection accuracy in that. SVM makes use of various hyper planes which specifies behaviour threshold for each attack. Depending on the classes of attacks available one can have ‘n’ number of hyper planes such that outliers those who does not fall into any of the hyper plane can be easily detected. SVM used for classification as well as regression. Proposed framework makes use of classification method to categorise packets into particular type of attack. As CICIDS dataset consist of numerical data for most of the features SVM shows efficient performance during training and testing phase of IDS evaluation.

3.4.4 Artificial Neural Network Classifier

Artificial Neural Network classifiers are mostly work as an computational node which take help of neural network such as biological networks in brains of animals. It allows working with a framework to help classifier to handle huge number of inputs. This is mostly used for anomaly based detection as these classifiers learns and act according to incidents happen before with a particular node. It tries to learn itself the normal behaviour of network and try to find some abnormal behaviour if any in network. Each packet is analysed with the help of features extracted from the dataset for a particular type of attack. Though the signature or previous data is not available still this classifier can sense the abnormal activity. This classifier provides highest accuracy in anomaly based detection as compare to other classifiers. It works with the number of input units with some features and all features will feed to the next layer as it is to analyse and classify attacks. Feed forward NN with back propagation is the popularly used algorithm from artificial intelligence domain.

4. EXPERIMENTAL RESULTS

Proposed framework is tested in real time environment. All the nodes are connected in the distributed environment and all classifiers will work together in ensembling to test real time traffic. All classifiers are trained using CICIDS dataset with reduced features after applying feature selection. Performance parameters consider for evaluation are detection rate, precision, false alarm rate and accuracy. All classifiers worked in both method of attack detection as signature based and anomaly based attack detection. Results show comparison between individual classifier with ensemble classifier. Performance parameters used for evaluation is derived from confusion matrix. Confusion matrix used to describe the instances of predictions generated by the classifier. Figure 2 shows the confusion matrix represented in the form of instances.

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Figure 2: Confusion Matrix

From Figure 2 we can observe the predicted and observed values used to evaluate the performance of IDS. These values are used to calculate accuracy, precision and many more parameters.

TP – Positive prediction value and the prediction is correct.

FP – Positive prediction value and the prediction is incorrect

TN – Negative prediction value and the prediction is correct

FN – Negative prediction value and the prediction is incorrect

All these values are used to calculate precision and accuracy. Formulas for calculating it is given in Equation (3) and (4) respectively.

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$accuracy = \frac{TP + TN}{total} \tag{4}$$

Where total – total number actual values

Another parameter false positive rate is also used for evaluation. FAR is the total number of wrong predictions made out of actual value. Detection rate parameter depends on the number of packets predicted correctly as compared to packets entered in the system. All the parameters are checked by testing classifier in both methods signature as well a anomaly based detection. Precision parameter shows the exactness of classifier which needs to be high with less false alarm rate. The more number of predictions wrong accuracy decrease for that classifier. Experimental results carried out for both methods as signature and anomaly based detection.

4.1 Performance Evaluation of Signature based detection

All the classifiers are trained using reduced features for CICIDS dataset obtained after applying feature reduction technique. Each classifier is trained and tested in real time environment for the mentioned evaluation parameters. All the results for individual and ensemble classifiers are recorded to provide comparison between them. All the classifiers are first trained for signature based attack detection where anomaly attacks will not be detected and passed to next detection method. Results recorded for all classifiers for signature based detection are represented in Table 2.

Table 2: Comparison of performance parameters using Signature based detection

Classifier	**RF**	**IF**	**SVM**	**AN N**	**Ensemble**
Precision	0.92	0.88	0.89	0.84	0.96
Detection rate	94%	90%	88%	87%	97%
False alarm rate	0.12	0.20	0.15	0.11	0.09
Accuracy	95%	94%	89%	86%	98%

From Table 2 we can observed that RF classifier provides good performance in precision and FAR. RF provides better accuracy as compared to other individual classifier. Classifier IF is mostly used for anomaly based detection so performance for this classifier when used in signature based method provides less performance as compared to RF. IF provides good accuracy and precision value, but more FAR as compare to other classifier. SVM and ANN provides less performance as compared to RF and IF classifiers. As SVM mostly work on binary data sometimes it provides biased output for prediction. That is the reason SVM shows less accuracy as compared to other classifiers.

Ensemble classifier shows highest performance as compare to individual classifiers. Accuracy obtained from ensemble method is 98% with highest precision value of 0.96, FAR reduced to 0.09 and obtained 97% detection rate. From table we can analyze that ensemble approach removes all limitations of individual classifier and provides good performance. Graphical represented of all parameters is shown in Figure 3.

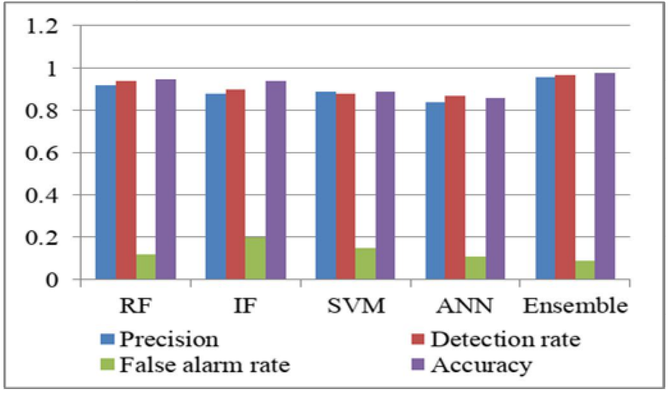

Figure 3: Graphical representation of performance parameters for signature based detection.

4.2 Performance Evaluation of Signature based detection

Packets which are not observed as attacks in signature based detection are allowed to enter in the network. Once such packets enters anomaly detection phase will start to check state of network for any abnormal activities. This phase is introduced in this framework to avoid limitation of signature based detection. To identify new attacks this phase is used as anomaly based detection. In this phase all classifiers are trained for normal behavior of network to identify attacks. This type of detection monitors the abnormal activity which deviates normal behavior of network. Performance of all classifier and ensemble method is recorded using anomaly based detection. Testing for this phase is also done in real time environment. Performance of all classifier along with ensemble approach is represented in Table 3.

Table 3: Comparison of performance parameters using Anomaly based detection

Classifier	RF	IF	SVM	ANN	Ensemble
Precision	0.88	0.92	0.78	0.84	0.92
Detection rate	95%	96%	88%	94%	97%
False alarm rate	0.25	0.15	0.28	0.17	0.13
Accuracy	93%	95%	89%	94%	96%

From Table 3 we can observed that new attacks can be easily identified by IF and ANN as both classifiers are learning on their own to find abnormal activities in network for attack detection. RF also provides good performance with 93%of accuracy but it is less as compared to other classifiers. SVM shows poor performance in anomaly based detection as compare to its performance in signature based detection. Ensemble approach shows awesome performance in this method as well with accuracy of 96%, 0.92 precision, FAR reduced to 0.13 and 97% detection rate. Graphical representation of performance of IDS using anomaly based detection is shown in Figure 4.

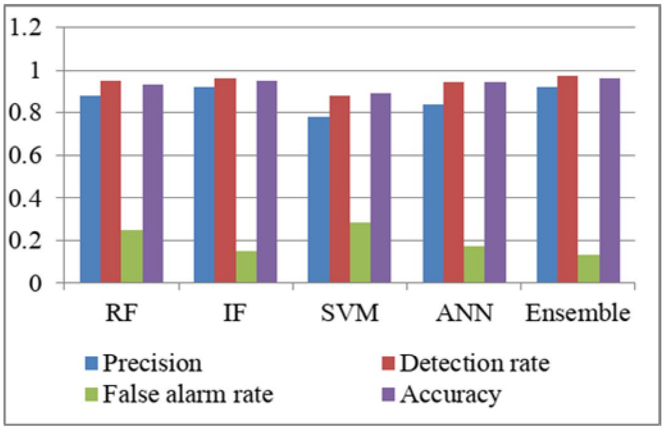

Figure 4: Graphical representation of performance parameters for anomaly based detection.

5. CONCLUSIONS

Proposed hybrid framework provides excellent performance in real time environment while working in distributed and ensemble approach. IDS system captures data in real time environment and analyze it using signature based method for attack detection. Dataset used for training and testing is CICIDS 2017. As the dataset deals with most modern attacks classifiers provides excellent performance in detection of real time modern up-to-date attacks. Packets which are observed as attacks in this method are rejected from network. Mostly up to 90% of attacks are detection using this dataset. Packets which are not detected as attacks in first phase are analyzed again using anomaly detection method to avoid network data loss. Anomaly based detection used to observe all packets passed by first phase of IDS for any abnormal activity. Sometimes new attacks cannot be detected by signature based detection due to lack of signatures in dataset. Such attacks can be detected in second phase with deviation of behavior. Attack detection in this phase avoids entry of intruders in network. Using this framework administrator can provide more security to data passing in the network and can analyze and detect new types of attacks as well. Experiments are conducted on four classifiers as RF, IF, SVM and ANN individually and in ensemble. All classifiers are trained for both types of methods to detect known as well as unknown attacks. Form experiments we can conclude that ensemble approach shows highest performance as compare to any individual classifier. However RF also provides good performance in case of both types of attack detection. IF classifier shows better performance for anomaly based attack detection as compared to signature based detection as this classifier can easily observe abnormal behavior of packets passing in network. SVM classifier shows average performance in both methods whereas ANN shows better performance in anomaly detection. Excellent performance is shown by Ensemble approach as compared to individual classifier with highest values in all performance evaluation parameters. In future this architecture can be extended further to save and create signatures of unknown attacks identified by anomaly based detection so that dataset can be updated with these signatures for improves performance of IDS in signature based attack detection.

REFERENCES

1. C. Guo, Y.Ping, N.Liu, S.Luo. **A two level hybrid approach for intrusion detection,** *Science Direct, Neurocomputing*, vol. 214, pp. 391-400, 2016. https://doi.org/10.1016/j.neucom.2016.06.021
2. G. P. Spathoulas and S. K. Katsikas. **Reducing false positives in intrusion detection systems**, *Science Direct, Computers and Security*, vol. 29, no. 1, pp. 35-44, 2010.
3. S.R.Khonde and V. Ulagamuthalvi. **Hybrid Architecture for Intrusion Detection System**, *Ingenierie des Systemes d'Information,* vol. 24, no. 1, pp. 19-28, February 2019.
4. P.Casas, J. Mazel, P. Owezarski. **Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge,** *Science Direct, Computer Communications*, vol. 35, no. 7, pp. 772-783, 2012.
5. H. Sarvari and M M. Keikha. **Improving the Accuracy of Intrusion Detection Systems by Using the Combination of Machine Learning Approaches**, *IEEE, International Conference of Soft Computing and Pattern Recognition*, 7-10 December 2010. https://doi.org/10.1109/SOCPAR.2010.5686163
6. S.R.Khonde and V. Ulagamuthalvi. **Ensemble-based semi-supervised learning approach for a distributed intrusion detection system,** *Journal of Cyber Security Technology, Taylor and Francis*, vol.3, no. 3, pp. 163-188, June 2019.
7. P. Aggarwala and S. Sharma. **Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection**, *Science Direct, Procedia Computer Science*, vol. 57, pp. 842-851, 2015.
8. S. Peddabachigaria, A. Abrahamb, C. Grosan, J. Thomas. **Modelling intrusion detection system using hybrid intelligent systems**, *Science Direct, Journal of Network and Computer Applications*, vol. 30, no.1, pp. 114-132, 2007.
9. S. Khonde and V. Ulagamuthalvi. **A Machine Learning Approach for Intrusion Detection using Ensemble Techniques - A survey**, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, pp. 328-338, 2018.
10. A. Shenfield, D. Day, A. Ayesh. **Intelligent intrusion detection systems using artificial neural networks**, *Science Direct, ICT express*, vol.4, no.2, pp. 95-99, 2018.
11. R. Ashfaq, X. Wang, J. Z. Huang, H. Abbas, Y. He. **Fuzziness based semi-supervised learning approach for intrusion detection system**, *Science Direct, Information Sciences*, vol. 378, pp. 484-497, 2017.
12. D. Barbara, N. Wu and S. Jajodia. **Detecting Novel Network Intrusions using Bayes Estimators**, *Proceedings of the 1st SIAM International Conference on Data Mining*, 5-7 April 2001. https://doi.org/10.1137/1.9781611972719.28
13. P. Rutravigneshwaran. **A Study of Intrusion Detection System using Efficient Data Mining Techniques**, *International Journal of Scientific Research in Network Security and Communication*, vol. 5,no. 6, pp. 5-8, 2017.
14. G. Giacinto, R. Perdisci, R. Mauro, R Fabio. **Intrusion detection in computer networks by a modular ensemble of one-class classifiers**, *Science Direct, Information Fusion*, vol. 9, no. 1, pp. 69-82, 2008.
15. A. Abdulla and I. Mamun. **A novel SVM-kNN-PSO ensemble method for intrusion detection system**, *Science Direct, Applied .Soft Computing*, vol. 38, pp. 360-372, 2016.
16. N. Littlestone Nick and M. K. Warmuth. **The weighted majority algorithm**, *Science Direct, Information and Computation*, vol. 108, no. 2, pp. 212-261, 1994.
17. S. Sindhu, S. Geetha, A. Kannan. **A. Decision tree based lightweight intrusion detection using a wrapper approach**, *Science Direct, Expert System and Applications*, vol. 39, no. 1, pp. 129-141, 2012. https://doi.org/10.1016/j.eswa.2011.06.013
18. R. Panigrahi and S. Borah. **A detailed analysis of CICIDS2017 dataset for designing Intrusion Detection Systems**, *International Journal of Engineering & Technology*, vol. 7, pp. 479-482, 2018.
19. T.S. Chou, K.K. Yen and J. Luo. **Network intrusion detection design using feature selection of soft computing paradigms**, *International Journal of Computational Intelligence*, vol. 4, no. 3, pp.196–208, 2008.
20. A. Suebsing and N. Hiransakolwong. **Euclidean-Based Feature Selection for Network Intrusion Detection**, *International Conference on Machine Learning and Computing, Montreal, IACSIT Press*. pp. 222–229, 2011.
21. S. Choi, H. Chae, B. Jo, T. Park. **Feature selection for intrusion detection using NSL-KDD**. *Recent Advances in Computer Science*, pp. 184-187, 2013.
22. S.R.Khonde and V. Ulagamuthalvi. **Fusion of Feature Selection and Random Forest for an**

Anomaly-Based Intrusion Detection System, *Journal of Computational and Theoretical Nanoscience*, **v**ol. 16, pp. 1–5, 2019.

23. S. Ikram and A. Cherukuri. **Intrusion detection model using fusion of chi-square feature selection and multi class SVM**, *Journal of King Saud University, Computer and Information Sciences*, vol. 29, no. 4, pp. 462-472, October 2017.
24. Y. Xiao, C. Xing, T. Zhang, Z. Zhao. **An Intrusion Detection Model Based on Feature Reduction and Convolutional Neural Networks,** *IEEE Access*, vol. 7, pp. 42210–42219, 2019. https://doi.org/10.1109/ACCESS.2019.2904620
25. N. Reddy, P. Vemuri, A. Govardhan. **An Implementation of Novel Feature Subset Selection Algorithm for IDS in Mobile Networks**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 5, pp. 2132-2141, 2019. https://doi.org/10.30534/ijatcse/2019/43852019
26. Maryam, N. Setiawan. **A Wrapper Feature Selection Based on Ensemble Learning Algorithm for High Dimensional Data,** *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 2782-2787, 2019. https://doi.org/10.30534/ijatcse/2019/16862019
27. G. Folino and P. Sabatino. **Ensemble based collaborative and distributed intrusion detection systems: A survey,** *Elsevier Journal of Network and Computer Applications*, vol.66, pp. 1-16, 2016.
28. Y. Zhou, G. Cheng, S. Jiang, M. Dai. **Building an efficient intrusion detection system based on feature selection and ensemble classifier,** *Computer Networks*, vol. 174, 2020.
29. S. Babu and M. Arasi. **A Hybrid Approach for Intrusion Detection using OPSO and Hybridization of Feed Forward Neural Network (FFNN) with Probabilistic Neural Network (PNN)-HFFPNN Classifier**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 206-201, 2020. https://doi.org/10.30534/ijatcse/2020/31912020
30. Hasan, Md., Nasser, Md., Ahmad, S. and Molla, K. **Feature selection for intrusion detection using random forest**, Journal *of Information Security*, vol. 7, no. 3, pp.129–140, 2016.
31. Hasan, M.A.M., Nasser, M., Pal, B. and Ahmad, S. **Support vector machine and random forest modeling for intrusion detection system (IDS)**, *Journal of Intelligent Learning Systems and Applications*, vol. 6, no. 1, pp.45–52, 2014..
32. X. Li, W. Chen, Q. Zhang, L. Wu. **Building Auto-Encoder Intrusion Detection System based on random forest feature selection**, *Computers and* Security, vol. 95, 2020. https://doi.org/10.1016/j.cose.2020.101851