International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 13, No.1 (Jan-2023)

http://dx.doi.org/10.12785/ijcds/130130

Ensemble of Decision Trees for Intrusion Detection System

D. P. Gaikwad[1] and D. Y. Dhande[2]

[1] *Department Computer Engineering, AISSMS College of Engineering, Pune, India*
[2] *Department Mechanical Engineering, AISSMS College of Engineering, Pune, India*

Received 23 Sep. 2021, Revised 16 Sep. 2022, Accepted 12 Jan. 2023, Published 31 Jan. 2023

Abstract: Now-a-days, Internet is playing vital role to change economic, political and social structure positively. In business transaction, the enormous assistance of Internet have stemmed in increased number of users and subsequently intruders. Intrusion Detection System detects intruders in networks. Using traditional approaches of intrusion detection, it is actual difficult to analyze packets in network. Development of Intrusion detection system by using ensemble method is leading to faster and enhance accurate detection rate. In this paper, decision trees based an ensemble classifier has proposed for detection of intrusion in network. The key aim of this research is to develop an ensemble to enhance accuracy of network intrusion detection on testing data set. Three decision trees have used as a base classifier. Because the decision trees are modest in environment and produce simple rules in if-then form. For building and testing the proposed ensemble based classifier, NSL-KDD dataset have used. The novelty of this research work is that ensemble of fast decision trees have combined together which provided very high accuracy. Experimental results shows that the proposed ensemble classifier beats its base classifiers and other existing ensemble classifiers on test dataset. It is also observed that the proposed ensemble classifier offers improved classification accuracy than Random forest and AdaBoost on test data-set. The proposed ensemble classifier also offers better accuracy than existing classifiers on training data-set. The proposed ensemble classifier also provide higher accuracy than classifier proposed in literature on 10-fold cross validation. Overall, the proposed ensemble based classifier beats standard ensemble classifiers and existing classifiers.

Keywords: Decision Trees, RepTree, Random Tree, J48, Ensemble, Precision

1. Introduction

Recently, Internet has become a vital communication media to reform our relations to daily events. The practices of Internet have been increased to exchange digital information across networks. It also promotes intruders to exploit faults in computers in different organizations by violating rules. It may damage computers, evidence on computers in organizations. For avoiding harmful effect, network security solution is needed to maintain integrity, confidentiality and availability [1]. To avoid these damages intrusion detection and other tools like firewall are available in market.Intrusion detection system is very important software is used to detect internal intruders. So, there is a need of developing fast and accurate IDS. Network Intrusion detection system is being used to protect information on computers in organizations. It also scans computers and network traffic to recognize and report violations. It helps in raising an alarm if any burglar activities happen in organization. This system uses network agents which observe and analyze network traffic. It also does inspection of packets in network to check its behavior. It is usually divided into three types: Anomaly, Signature grounded intrusion detection and analysis of State-full protocol [2]. Anomaly grounded IDS use statically way to detect unknown attacks. It detects activities which are deviates from normal behaviors in network. Anomaly based IDS gives high false negatives and positive negative rates. Signature based IDS do not recognize unknown attacks. It uses signature of recognized attacks which are available in database. State-full protocol analysis relates documented protocol profiles to identify a random sequence of commands in both application layer and networks. Each detection type has advantages and drawbacks. For recognizing unknown attacks, anomaly-based IDS tool which is widely used in computer network [3]. Some disadvantages lead to false negatives, false positives, increases CPU usages and slow networks. In literature survey, it is observed that machine learning techniques offers high detection rate [4]. For overcoming limitation of present intrusion detection, modern and old machine learning approaches are being widely used. Decision tree, Support Vector Machine, Bayesian Classification and ANN are usually used in IDS field [5]. A single classifier is not

E-mail address: dpgaikwad@aissmscoe.com

skilled to detect each kind of attack. For addressing these issues, some authors have addressed and ensemble classifier for IDS. Recently, Ensemble classifiers play energetic role in increasing detection accuracy of IDS. Ensemble scheme is a combination of distinct classifiers which outperforms individual classifiers [6]. It is very difficult to select appropriate base classifier in ensemble method. The main goal of this paper is to select suitable individual classifiers of ensemble classifier to increase classification accuracy, decrease rate of false positive on training dataset with less training time. Specifically, decision trees take less time for training. Decision tree represent as if-then rules set and is simplest machine learning approaches [7]. Ensemble of decision trees has increased accuracy of IDS. In this research paper, a new ensemble classifier structure has proposed for ID system. Three different decision trees have combined to construct ensemble classifier.In this research, decision trees are used as base classifiers which are fast and produce simple rules. Due to these base classifiers, the proposed ensemble classifier take lesser time to build than existing ensemble classifiers using heterogeneous base classifiers. The proposed ensemble classifier is useful to implement real time intrusion detection system which takes very less time for packets processing. For experimental study, KDD-99 dataset have used for training individual and proposed ensemble based classifier. Overall, the paper is planned in different sections. In section 2, some research work has been discussed. Section 3 used for discussion on the offered ensemble based classifier. In section 4, experiment based results have explained. Finally, conclusions and future scope are given in section.

2. Literature Survey

In this section, some existing ID system have discussed and analyzed. Harek Haugerud at.el, [8] have proposed an elastic parallel network IDS. This system is built on rule distribution and NFV prototype. They have developed two adaptive algorithms which adjust, divide and orders signatures rules dynamically. Algorithms are capable to split work load of IDS which enable to scale system. Investigational results display that devised algorithms must tune to avoid some packets drives unexamined. Pushparaj Nimbalkar and Deepak Kshirsagar [9] have proposed IDS based on attribute selection method. They have used Info Gain and Gain Ration method witch select top 50 features from training dataset. This system have evaluated on Cup 1999 and IoT-Bot datasets using JRip classifier. Investigational results display that proposed system offers advanced performance than original dataset. Mahmoud Said at.el, [10] have offered a new DL-based hybrid ID system. This hybrid system is grounded on CNN. Authors have offered a SD-Reg novel regularizing technique based on weight matrix's standard deviation. Results display that this regularizing is useful to evade over fitting and increase capability of Intrusion detection. Pooja T. S [11] has developed an automated technique for network based IDS. UNSW-NB15 complex and KDDCUP-99 datasets have used for training system. The Long Short Term Memory deep learning has trained. LSTM offers accuracy of 99 on both datasets. They have suggested intrusion detection using Convolution Neural network. Shaohua L.V [12] has introduced IDS using recurrent neural networks. This system deals with long sequential problems which introduce the sequence-to-sequence model. This system attained sound prediction performance using ADFA-LD test data set. F. J. Mora-Gimeno [13] has proposed deep neural network based IDS. It adds multiple detection methods using call graphs. The integrated model offers shows the advanced detection values and lesser false positives than individual techniques. It offers a success rate of 98.8 for complex datasets and 100 for simple datasets. Tian Xinguang at.el, [14] have proposed host based intrusion detection scheme which monitor system call activities. For characterizing normal behavior, a Markov chain homogeneous model has used which associates unique calls with states of Markov chain. The proposed online detection method gives care to both accuracy and training efficiency. Guo Pu et.al, [15] have proposed unsupervised machine learning to develop and implement IDS. This offered system is a combination of Sub-Space Clustering and Support Vector Machine which has a skill to detect strange attacks without any previous information. Authors have used NSL-KDD dataset to assess the offered network intrusion detection scheme. R. M. Gomathi and M.Nithya [16] have proposed defense-in-depth IDS. They have incorporated processes of attack analytical procedure for IDS System uses a reconfigurable digital networking policy to recognize and battle Virtual Machine zombie attempts. Raymond Mogg at.el, [17] have proposed an intrusion detection tool that is based on Decision Tree. NSLKDD dataset have used for training Decision trees. In this proposal, Genetic algorithm select relevant feature and find the feasibility of producing understandable dodging attacks against IDSes. Experimental results offered attacks that like to a given seed attack are classified as benign for both the teardrop and Nmap attack types. J. Olamantanmi Mebawondu at.el, [18] have proposed Artificial neural network based IDS. Authors have used UNSW-NB15 dataset for training ANN. Continuous attributes have discretized in binary before training ANN. The offered model shows optimistic correlation value 0.57 and gives class accuracy of 76.96. Ahmed Mahfouz at.el, [19] have offered ensemble based IDS using GTCS dataset. This system has overcome the faults of some the existing available datasets. M. A. Jabbar ET. Al, [20] have suggested an ensemble classifier for IDS using Random Forest and Average One-Dependence Estimator. They have detected that Random forest improved accuracy with less error rate. The offered ensemble classifier offers 90.52 accuracy with FP rate 0.14.

3. METHODOLOGY OF THE PROPOSED ENSEMBLE CLASSIFIER

A. Preparation of Data set

In this paper, NSL-KDD dataset have utilized for training and testing individual base classifiers and the offered ensemble based classifier. NSL-KDD dataset is publically available and widely adopted by researchers for IDS. This dataset has desired samples of the complete KDD data set. This dataset is developed form of KDD Cup99 Dataset in which redundant samples have removed to prevent biased result [21]. NSL-KDD dataset includes 42 features with a class label attribute. The available NSL-KDD dataset have pre-processed to refine training dataset and testing datasets. This refined training dataset comprises of 67,343 normal samples and 58,638 anomaly samples. The dimension of training dataset is 125,981 samples. This refined testing dataset consists of 9,711 normal samples and 12,833 anomaly samples. The size of testing dataset is 22,544 samples. These datasets are utilized to train base classifiers. Initially, Decision Tree J48 have trained and verified using refined NSL-KDD dataset. After training J48 DT REPTree and Random Tree have trained and tested using same dataset. These three decision trees have combined using Average of Probability combination rules. Ensemble classifier verified using refined NSL-KDD dataset.

B. Introduction to base classifiers

The main goal of this investigation is to implement a new ensemble classifier for IDS. Selections of basic classifiers are very essential to develop ensemble classifier. Random Tree, J48 and RepTree have used as a base classifier. In this section, J48, Random Tree, RepTree and the offered ensemble based classifier have discussed.

C. J48 Decision Tree

The decision tree does sorting process through symbols of nodes and branches. Attributes are indicated by nodes and splitting of the attributes is denoted by branches. In decision tree, leaves of tree denote classes of dependant variable. Level of node in tree depends on information gain ratio of attribute. Each attribute node is selected for further branching. Selected node is split according to info gain of attribute. J48 DT classifier can forecast the class label of a test sample in a dataset from list of independent and dependent variables. J48 classifier performs the pruning of the tree. It is capable to handle classification with the absent values in data and handles both discrete and continuous variables. It reduces the error rate by replacing internal nodes with a leaf node and manages high dimensional data [22]. Algorithm 1 describes the procedure of construction of J48 decision tress.

Algorithm 1 : Construction of J48 Decision Tree.
Step 1: Calculate Entropy (D); D is training dataset

$$Entropy(D) = \sum_{i=1}^{\infty}(\frac{Frequency(C_i, D)}{|D|})\log_2(\frac{Frequency(C_i, D)}{|D|})$$

Where D dimension of D, Ci is Dependent variable, N is Class's number and frequency (Ci, D) is the samples included in class Ci.
Step 2: Calculate the InfoGainx (D) of X test attribute

$$InforGain_x(D) = Entropy(D) - \sum_{i=1}^{L}(\frac{|D|}{|D|})Entropy(D_i))$$

Where L: test outputs X, Di is a subclass of D corresponding to ith output,
Step 3: Calculate Split Info(X) obtaining for D partitioned into L subsets.

$$SplitInfo(X) = -\sum_{i=1}^{L}(\frac{|D_i|}{|D|})\log_2(\frac{|D_i|}{|D|})+(1-(\frac{|D_i|}{|D|})\log_2(1-\frac{|D_i|}{|D|}))$$

Step 4: Calculate the Gain Ratio(X)

$$GainRatio(X)) = \frac{InforGain_x(D)}{InforGain(X)}$$

Step 5: Highest gain attribute is elected as the root node. Repeat Step 1 to step 4 for every middle node until all the examples are reaches the leaf node.

D. Random Tree and REP-Tree

This tree work J48 tree which select attributes randomly and it do not perform pruning. REP-Tree is grounded on C4.5 algorithm and it takes very less time to build. REP-Tree yield classification on discrete values and regression tree on continuous values using variance, information gain. It prunes tree using back fitting and reduced-error pruning technique [23]. These three base classifiers can be combined using different combination rules. Majority Vote and Average of probability combination rules are widely used in classification. It is found that for this application Average of probability combination rules is suitable combination rule method. In following equation, class C*is the value of weighted majority vote.

$$C^* = argmax_{c\epsilon C}(\sum_{e\epsilon E)} r_e d_{e,c})$$

Where re is a reliability weight for classifier e and

$$d_{e,c} = \begin{bmatrix} 1; & if & e & outputs & cand0; & otherwise \end{bmatrix}$$

In this research, average of class probability combination rule has used to combine three decision trees. Figure 1 depicts the steps of combination of three decision trees. Symbol indicates the combination formula.

4. PROPOSED ENSEMBLE BASED CLASSIFIER FOR INTRUSION DETECTION SYSTEM

In this section, the architecture of the offered Ensemble classifier has discussed. As traditional intrusion detection system has some draw-backs. Some system cannot detect strange attacks and has poor adaptability. Most of the present system take more model building time and produce high false positive rates. This research goal is to plan and develop an ensemble classifier with suitable base classifiers. For this purpose, simple decision trees have used to suggest an ensemble classifier. Three decision trees, J48, Random

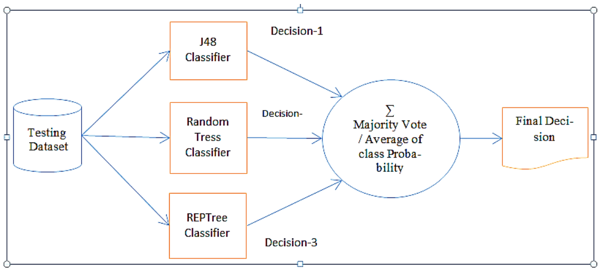

Figure 1. Ensemble of Decision Trees Using Average of Class Probability

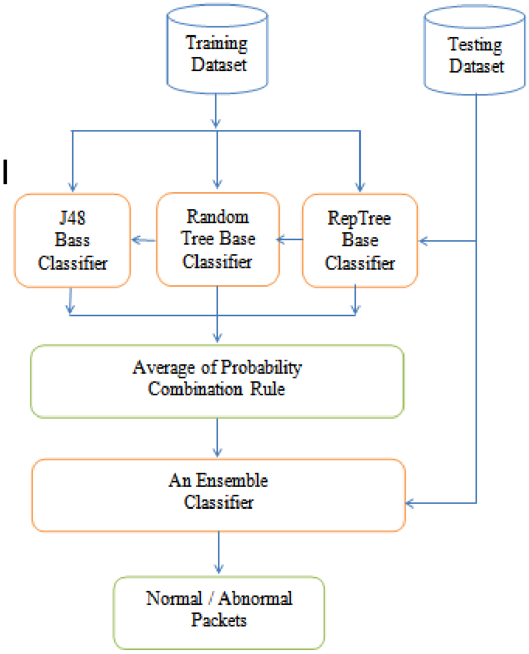

Figure 2. Architecture of the suggested IDS

decision tree and REP Tree have deliberated as base classifiers because these are actual simple in implementation and they take very fewer times to build. Initially, three decision trees based base classifiers have trained and analysed on test data set. The performances of classifiers have measured in terms of model building, classification accuracy, precision and recall. In second stage, base classifiers have combined using average probability combination rule. In Figure 2, the recommended architecture of IDS has shown. As presented in Figure 3, refined NSL-KDD data set used for evaluating base classifiers. These trained and tested final base classifiers joined using combination rule to create an ensemble classifier. The offered an ensemble trained using same refined data set. Finally, an ensemble classifier tested on test data set. Final Ensemble classifier saved for recognition of normal and unusual packets in network.

5. EXPERIMENTAL RESULT ANALYSIS

This section carries out Experiment based results analysis using following metrics. Most performance measures used like the model building time, precision, accuracy and recall to measure the proposed ensemble and base classifiers. The following equations [1-2] are used to compute different measures.

$$Accuracy = ((TP + TN))/((TP + TN + FP + FN)) \quad (1)$$

$$Precision = TP/((TP + FP)) \quad (2)$$

$$Recall = TP/((TP + FN)) \quad (3)$$

Where, TN is true negative and TP is true positive, FN is false negative and FP is false positive. Every ensemble classifier takes more building time because it combines all results of base classifiers. Basically, accuracy on test data set is very important aspect for any classifier. In table 1, the performances of base classifiers and the offered ensemble classifier have listed in term of accuracy and false positive rates on test data-set. In table 2, the ability of all basic classifiers and the suggested ensemble classifier have given in term of model building time, precision and recall value on test data set. Basically, accuracy on test unknown sample is very important for any intrusion detection system. The proposed ensemble classifier gives less accuracy than Random forest and AdaBoost on training dataset, but proposed ensemble classifier gives more accuracy on test dataset than Random forest and AdaBoost.

In Figure 3, accuracy of classification and false positive rates of base classifiers and proposed ensemble based classifier on testing data-set have shown. From Table 1 and Figure 3, it can be determined that the proposed ensemble offered better accuracy than its base classifiers on test data-set. It also can be observed that the suggested ensemble based classifier gives better accuracy than AdaBoost and Random Forest ensemble classifiers on test data-set. The proposed classifier also provides better accuracy of classification than classifier proposed in Ref.3 and Ref.18 using test data-set. From Table 2, it can also be detected that it provides better precision value than classifier proposed in Ref.18 with smallest false positive rate and best recall value.

In Table 3, the abilities of base classifiers and the offered ensemble classifier have listed in term of class accuracy and false positive rates on training data-set. In Table 4, the performances of basic classifiers and the offered ensemble based classifier have listed in term of model building time, precision and recall value on training data-set. In figure 4, false positive rates and accuracy and offered ensemble classifier on training data-set have shown. According to Table 3 and Figure 4, ensemble classifier offered better classification accuracy than its two base classifiers on

TABLE I. ACCURACY AND FALSE POSITIVE RATE OF CLASSIFIERS ON TEST DATASET

Sr. No.	Model	Test Accuracy in %	FP
1	Proposed ensemble based classifier	82.0263	0.155
2	J48	81.5339	0.146
3	Random Tree	81.3565	0.160
4	RepTree	81.5073	0.162
5	Random Forest	80.4516	0.155
6	AdaBoost (J48)	77.8522	0.175
7	Changjian Lin at.el [in Ref.3]	72.9800	0.663
8	J. Olamantanmi [in Ref.18]	76.9600	0.206

b

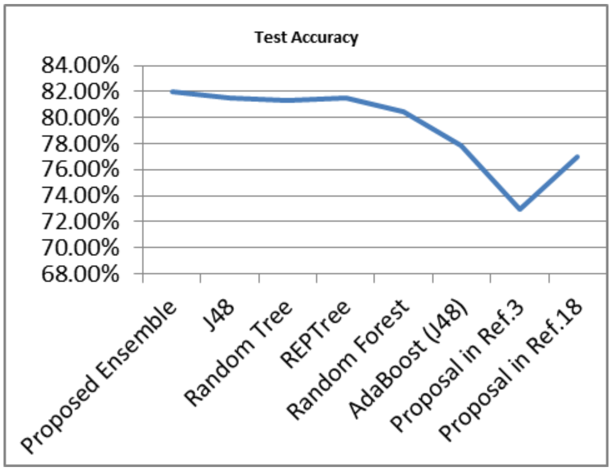

Figure 3. Accuracy and FP of Classifiers on Test Dataset

TABLE II. MODEL BUILDING TIME, PRECISION, RECALL VALUES OF CLASSIFIERS ON TEST DATASET.

Sr. No.	Model	Model Building Time	Precision	Recall
1	Proposed ensemble based classifier	38.18 seconds	0.842	0.820
2	J48	0.94 seconds	0.858	0.815
3	Random Tree	NA	0.837	0.814
4.	RepTree	0.5 seconds	0.835	0.815
5	Random Forest	1.39 seconds	0.852	0.805
6	AdaBoost (J48)	264.04 seconds	0.838	0.779
7	Changjian Lin at.el. [in Ref.3]	Not Given	NA	NA
8	J. Olamantanmi [in Ref.18]	NA	0.798	NA

training data-set. The offered ensemble classifier offers less classification accuracy than AdaBoot, Random forest and AdaBoost ensemble classifiers training data-set. It is also determined that the offered ensemble based classifier gives improved accuracy than classifier proposed in Ref.4 on training data-set. It is also observed that all classifiers offer same false positives rates on training data-set. According to table 4, the proposed ensemble took additional time to train than its base classifiers. Precision and recall values are almost for all classifiers on training data-set.

In Table 5, the ability actions of base classifiers and the suggested ensemble basic classifier have listed in term of classification accuracy and FP rates on cross validation. In Table 6, the capability of base classifiers and the suggested ensemble based classifier have listed in term of model building time, precision and recall value on cross validation. In figure 5, false positive rates and accuracy of base classifiers and proposed ensemble based classifier on 10-fold cross-validation have shown. According to Table 5 and Figure 5, it observed that the suggested ensemble classifier offered well classification accuracy than its all base classifiers on cross validation. The offered ensemble classifier offers less classification accuracy than Random forest and AdaBoot ensemble classifiers on CV. It also can be determined that the offered ensemble based classifier

TABLE III. ACCURACY AND FALSE POSITIVE RATE OF CLASSIFIERS ON TRAINING DATASET.

Sr. No.	Classifiers	Training Accuracy in %	FP
1	Proposed ensemble based classifier	99.9603	0.000
2	J48	99.9111	0.001
3	Random Tree	99.9944	0.000
4	RepTree	99.8825	0.001
5	Random Forest	99.9944	0.000
6	AdaBoost(J48)	99.9944	0.000
7	Pooja T. S Ref.[4]	99.7000	NA

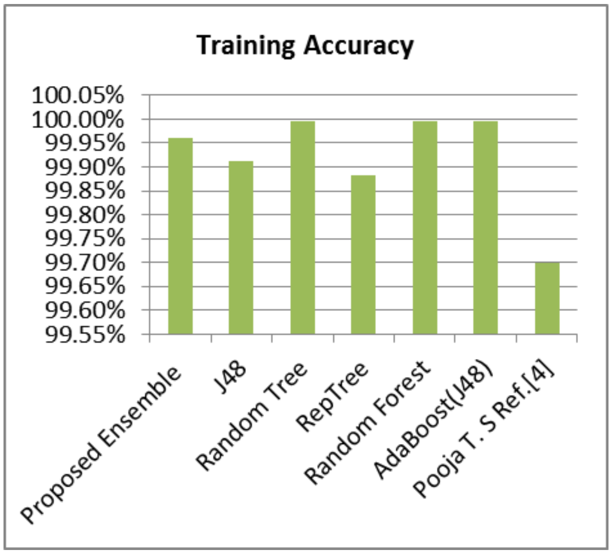

Figure 4. Accuracy and False Positive of Classifiers on Training Dataset

TABLE IV. MODEL BUILDING TIME, PRECISION, RECALL VALUES OF CLASSIFIERS ON TRAINING DATASET.

Sr. No.	Model	Model Building Time	Precision	Recall
1	Proposed Ensemble	42.13 seconds	1.000	1.000
2	J48	0.55 seconds	0.999	0.999
3	Random Tree	1.58 seconds	1.000	1.000
4	RepTree	1.14 seconds	0.999	0.999
5	Random Forest	4.95 seconds	1.000	1.000
6	AdaBoost(J48)	0.58 seconds	1.000	1.000
7	Pooja T. S Ref.[4]	NA	NA	NA

gives much better accuracy than classifier offered in Ref.20 on10-fold cross validation. It is also detected that the offered ensemble classifiers offer very less false positives rates than its base classifiers and proposed classifier in Ref. 20 on cross validation. According to Table 6, the proposed ensemble takes more time to train than its base classifiers. Precision and recalls of the offered ensemble classifier are better than its basic classifier but equal to values of Random forest and AdaBoost ensemble classifiers.

6. Conclusions and Future Work

In this paper, a novel ensemble classifier has proposed for Intrusion detection system. Three decision trees have used as a base classifiers. J48, REP-Tree and Random tree have utilized as a base classifiers of ensemble classifier. All base classifiers have combined using average of class probability to implement the proposed ensemble based classifier. Refined NSL dataset have utilized for training and testing all base and ensemble classifier. The proposed ensemble classifier have compared with its base classifiers and other two standard existing ensemble classifiers. Experimental results show that the proposed ensemble based classifier outperforms all it's base classifiers and existing classifiers. The proposed ensemble classifier provides improved classification accuracy than Random forest and AdaBoost on

TABLE V. ACCURACY AND FALSE POSITIVE RATE OF CLASSIFIERS ON CROSS VALIDATION.

Sr. No.	Model	Cross Validation Accuracy in %	FP
1	Proposed ensemble based classifier	99.8873	0.001
2	J48	99.7817	0.002
3	Random Tree	99.7658	0.002
4	RepTree	99.8357	0.002
5	Random Forest	99.9174	0.001
6	AdaBoost(J48)	99.9047	0.001
7	M. A. Jabbar at.el, [ref. 20]	90.5100	0.140

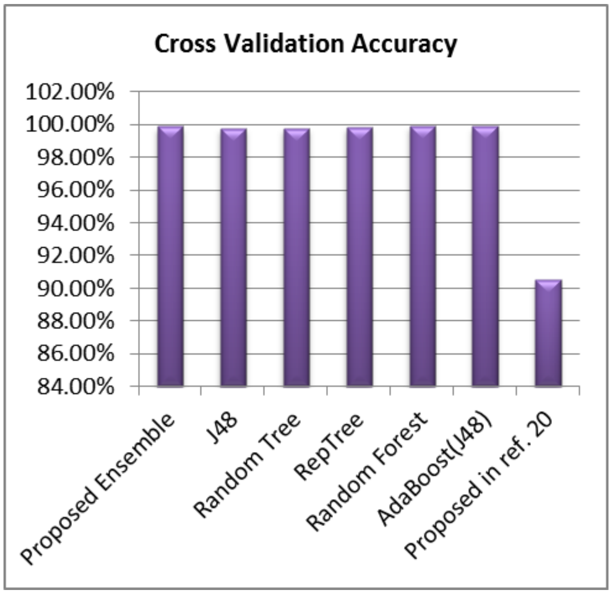

Figure 5. Accuracy and FP of Classifiers on cross validation

TABLE VI. MODEL BUILDING TIME, PRECISION, RECALL VALUES OF CLASSIFIERS ON CROSS VALIDATION.

Sr. No.	Model	Model Building Time	Precision	Recall
1	Proposed ensemble based classifier	39.87 seconds	0.999	0.999
2	J48	34.49 seconds	0.998	0.998
3	Random Tree	1.93 seconds	0.998	0.998
4	RepTree	6.36 seconds	0.998	0.998
5	Random Forest	85.83 seconds	0.999	0.999
6	AdaBoost(J48)	261.57 seconds	0.999	0.999
7	M. A. Jabbar at.el, [ref. 20]	NA	NA	NA

test dataset. It gives better accuracy than existing classifiers on training dataset and on cross validation. Overall, the proposed ensemble based classifier outperforms its all base classifiers and standard available ensemble classifiers. This system can be used to implement real time intrusion detection system.

Acknowledgment

We would like to express our deepest appreciation to to our Hon. Secretory, Maloji Raje for providing research facilities AISSSM College of Engineering. We would like to extend my sincere thanks to our mentor Dr. M.B Kokare for technical support to complete this research work. We had the pleasure of working with all colleagues.

References

Dr D P Gaikwad has completed his B.E (Computer Science and Engineering) from SGGS College of Engineering, Nanded in 1995. He has completed his M. Tech. (Computer Science and Engineering) in 2006 from College of Engineering, Pune. He has completed his Ph.D (Computer Science and Engineering) in 2017. He is working as Associate Professor and Head of Computer Engineering Department since 2013. He has published more 40 papers in international journals and conferences. He is a reviewer of international journal by Springer, IEEE Access and Elsevier.

Dr D Y Dhande is working as Associate Professor in AISSMS College of Engineering, Pune. He completed his PhD from COEP, Pune and has 21 years of teaching experience. He is reviewer of many reputed journals and published many papers. His areas of expertise include Tribology, Computational Fluid dynamics of bearings and fluid flows, Hydrodynamic Journal bearing analysis, bio fuels and structural analysis using finite element analysis.