# Intrusion Detection System Using Ensemble of Decision Trees and Genetic Search Algorithm as a Feature Selector

D. P. Gaikwad

Department of Computer Engineering, AISSMS College of Engineering, Pune Maharashtra, India
Tel: +91-9822609276 e-mail: *dpgaikwad@aissmscoe.com*

ORCID iD: 0000-0002-5014-1982

Research Paper Received: 17.09.2019 Revised: 22.04.2020 Accepted: 01.06.2020

**Abstract**—In the middle of this wonderful Internet technology, the rise and growth of Internet misuse is shocking which compromises the security of the computers in network. In doing so, the use of the Internet becomes very destructive for one and all. Any unauthorized person can steal private information by hacking computer. Anonymous attack has many causes, such as viruses, malware, misuse of privileges on the computer and unauthorized access to information systems. To reduce the exposure to such types of threats, organizations need a reliable, robust and fast computer network security mechanism. Intrusion detection is a mechanism which detects and prevents different intruders in internet. There are many techniques of machine learning which can be used to apply intrusion detection systems. Currently, many researcher are using ensemble method of machine learning to implement intrusion detection system. In ensemble method, the selection of suitable base classifiers is a very key process.

This paper propose a novel intrusion detection systems using ensemble of two well-known decision trees. C4.5 decision tree and Random Forest have selected as a base classifiers. Intrusion detection system is framed by combining the gains of both C4.5 and Random Forest decision trees. The performance of the proposed ensemble for intrusion detection system has evaluated in terms of classification accuracy, true positives and false positives. The experimental results show that the proposed ensemble classifier for intrusion detection performs well in classification accuracy, true positive than individual decision trees on testing dataset. Other aspects of performance of classifiers are described in the paper.

**Keywords**—Ensemble classifier, Random Forest, C4.5, Classification Accuracy, False Positives

## 1. Introduction

Recently, computer systems in network come across malicious code and intruders. These malicious and invaders can access, operate and disable computer systems across the Internet. To prevent these activities in Internet, there is need of a strong security mechanism. For detecting and preventing internal attacks in computers within network require strong system. Intrusion Detection System is powerful tool to detect internal attacks that illegally access computers of social group. It detects malicious activities by analysing the packets to prevent damage from the attack. Anomaly and misuse detection techniques normally are used for intruder detection. Anomaly intruder detection technique is a profile centred which examines normal traffic. It really detects the unknown packets in the network, but they are not so forceful in the detection rate. They also provide high false positive rates. Selection of suitable method and technology of IDS is exciting task to reduce the false positive rate. Machine learning and data mining techniques are commonly used in detection of intrusion. Although, such techniques are well innovative, there still arises a need to devise a proper technique and method to implement intrusion detection system. To overcome disadvantages of anomaly detection technique of intrusion detection, ensemble approaches of machine learning are being used by many researchers. Soft computing techniques are also used for deceasing false positive rates. Ensemble method of machine learning is more efficient which is more effective to decrease false alarms and elevate classification accuracy. Boosting and Bagging ensemble methods are mostly used to implement intrusion detection system. Stacking ensemble method is not so suitable for intrusion detection system because weak classifiers of it require more time to train [1] [2]. Different base classifiers can be combined together using different combination rules to improve the classification accuracy. You can combine any number of base classifiers from different categories of classifiers. In this paper, an intrusion detection system has proposed using combination of two decision trees. C4.5 and Random forest decision trees have used as base classifiers of ensemble classifier. The selection of correlated features from the dataset is very important to improve model building time and classification accuracy with low false positives. The significant features are selected based on their importance to identify the types of attacks. Genetic algorithm has used to choose relevant attributes from National Science Laboratory KDD training and test dataset. Finally, the performance of proposed system has assessed in term of accuracy, true and false positive rates. In rest of the paper, Section 2 is used to discuss current trends of intrusion detection system. In section 3, base classifiers are introduced. Section 4 describes proposed ensemble classifiers in detail. In Section 5, experimental results are discussed and analysed. Finally, Section 6 is dedicated for conclusion.

## 2. Literature Survey

In this section, existing intrusion detection systems have reviewed in detail. Many researchers are taking to implement accurate intrusion detection systems using different techniques and methods. In this survey, some significant proposals have discussed below. Haipeng Yao, Danyang Fu, Peiying Zhang, Maozhen Li, and Yunjie Liu [3] have proposed a multilevel intrusion detection system which is named as multilevel semi-supervised machine learning (MSML). This proposed system addresses two factors. One is imbalance of network traffic in different categories. Other is the non-identical distribution between training set and test set in feature space. This framework includes four modules. In pure module, they have proposed a hierarchical semi-supervised k-means algorithm to find out all the pure clusters. In the pattern discovery module, authors apply cluster-based method for finding unknown patterns. Then a test example is sentenced to label unlabelled unknown or known pattern. The FC unit can achieve FC for those unknown pattern samples. The model updating unit provides a mechanism for retraining. The proposed intrusion detection system can successfully differentiate known pattern samples and unknown pattern samples from the entire dataset. The known pattern samples are ensured high accuracy of 99.3%. This proposed system has some limitations. Authors have suggested overcoming these limitations. Yihan Xiiao, Cheng Xing, Tanining Zhang and Zhongkai Zhao [4] have proposed network-basedintrusion detection model based on a convolutional neural network. Using different dimensionality reduction methods, the redundant and irrelevant features in the network traffic data are removed. CNN have used to automatically extract features. Original traffic is converted into image format to reduce the computational cost. The experimental results specify that the proposed CNN IDS model van be used to improve the classification detection performance and significantly reduces the classification time. Proposed model can satisfy the real-time requirements of the intrusion detection system. Wajdi Alhakami et. al., [5] have developed a novel fully Bayesian-based approach for unbounded Generalized Gaussian mixture model. It integrates a feature selection mechanism to prevent irrelevant features. Bayesian inference methodology is stimulated to avoid under and overfitting. Infinite assumption has capability in learning simultaneously the model's parameters and number of components. The efficiency of proposed structure is confirmed by testing it on anomaly intrusion detection. Future work of authors is to offer large scale IDS-based datasets to propose a deep full analysis and detection system. Jia Jinfping, Chen Kehui, Chen Jia, Zhouu Dengwen and Ma Wei [6] have used Support Vector machine and RF classifiers to detect and recognize network evasions by means of intra and inter packet behaviour classification. By applying evasion technique, they have formed a network evasion trace set on normal TCP streams. A network evasion dataset built by converting TCP streams to codeword streams. The statistical features have extracted from the intra and inter packet. Using this dataset, SVM and RF classifiers trained to distinguish unlike types of evasions from normal streams. Authors have observed an accuracy of 98.95%.Peng Wei, Yufeng Li, Zhen Zhang, Tao Hu, Ziyoung Li and Diyang Liu [7] have proposed deep belief network for intrusion detection system. Optimization of deep belief network is essential for intrusion detection system. Authors have proposed a new combined optimization algorithm to optimize the deep belief networks. Initially, particle swarm optimization is designed based on the adaptive inertia weight and learning factor. Secondly, authors have used the fish swarm behaviour of cluster to optimize the PSO. Then, genetic operators are used to optimize the PSO to search the global optimization solution. Using this joint optimization algorithm, the global optimization solution is used as the network structure for intrusion detection classification model. The experimental results show that proposed algorithm shortens the average detection time. Sengupta [8] have used reinforcement Q-learning and RST to implement intrusion detection system. The dimensionality of training has reduced to increase accuracy with fewer discredited features. Kim et.al, [9] use hierarchical model using methods of anomaly and misuse detection. In this paper hybrid method have used. For misuse detection they have used C4.5 classifier. The one-class SVM model is used to reduce rate of false positive. Feng et. al. [10] have implemented intrusion detection system using SVM and self-organized ant colony network base clustering. Kuanga et.al, [11] have used hybrid approach for intrusion detection system. They have used Genetic algorithm and SVM to propose system named hybrid KPCA-SVM. KPCA-SVM is used to abstract the key feature of intrusion detection system. The Support vector machine multilayer classifier is used to identify an attack. Ezzat et.al, [12] have used C4.5, Naïve and Neural Network to implement intrusion detection system. It is multilayer intrusion detection model. It is found that classification accuracy of C4.5 decision tree is very grate as compare to Naïve Bayes and Multilayer. The planed model detects attack in first phase and in second phase it classifies attack. Zander et.al, [13] used unsupervised machine learning technique to classify traffic and classify applications in network.

Technique of feature selection has used to select optimal set of flow attributes. They determined the effect of dissimilar attributes. Creech and Hu [14] they have increased detection rates by using a new host-based anomaly based discontinuous system call patterns. They have created ADFA Linux data set using modern contemporary hacking methods and operating system. Muamer et.al, [15] applied expert system and data mining to implement intrusion detection. The system is developed in WEKA machine learning tool. They found that the performance, efficiency of detection and false positive rate are better than the current system. Hu and Yu [16] have used improved incremental HMM stochastic procedure to implement system call-based IDS. The pre-processing of dataset approach is used to speed up a hidden Markov model. Panda et.al, [17] have used supervised and unsupervised classifiers for data filtering in intrusion detection system. The system is hybrid of both learning. The output of first classifier is applied to second classifier for classifying the training and testing dataset. Due to this, overall performance is improved and the system made very intelligent in decision. Chung et.al, [18] have implemented intrusion detection system for cloud of networks. The proposed system is graph based model called as NICE system. The NICE is multiphase distributed vulnerability detection system to prevent virtual machines in cloud from being compromised. It provides reconfigurable virtual network-based solution. The NICE specifically is used to detect zombie exploration attacks in Infrastructure-as-a-Service (IaaS) clouds.

## 3. C4.5 and Random Forest Decision Trees Base Classifiers

Decision trees are predictive models which are used in machine learning, statistics and data mining. In data mining, decision trees are very popular and widely used for several problems. They are very popular because they are logic based and easily interpretable. Decision trees are the combination of mathematical and computational techniques which can used to describe and generalize dataset. These decision trees based models map observations about an item to conclude target value of item. Items are represented in branches and target values are presented in leaves of tree. In tree model, if there are finite set of values of target variable then these models are called as classification tree. In tree structure model, class labels are presented by leaves and conjunctions of features are represented by branches. In decision trees, if target variable take continuous value then such trees are called as regression tree. It is used to describe data and can represent decisions. In decision analysis, it can be used to make decisions. Decision activity can be traced as a sequence of simple decisions. In decision tree based classifier, a knowledge base is captured in a well-organized manner [19].

### 3.1. C4.5 Classification tree

C4.5 decision tree is developed by Ross Quinlan which is an improved version of ID3. The algorithm provides very decent performance and it is very simply to understand. C4.5 is best suitable classifier for many classification based applications. C4.5 algorithm builds a tree as a learning model from the data samples [20] [21]. In this algorithm, the divide and conquer scheme is used to build a decision tree. In divide and conquer scheme, partition of data is performed until every leaf cover cases of a single class. For avoiding over fitting, C4.5 algorithm prevents subdivision of some sets of training cases using some stopping criterion. It removes some of structures of produced decision trees to avoid over fitting. Using recursive divide and conquer strategy, C4.5 decision induction arises with training dataset which is divided at every node ensuring in smaller partitions. Every sample has related with a class label that identifies whether sample fits to a particular class or not. If any sample fall indifferent class, then splitting can be further performed. Greedy method is used for partitioning training dataset. Attribute selection measures are used to select type of splitting that occurs on a node. Attribute selection measures such as Gini Index, Information gain are used to partition a node. Gini Index measure is use to partition a node into a binary and Information gain divide a node into multiple ways. In ID3, information gain attribute selection measure is used to produce n-ary tree in some cases. Therefore, ID3 cannot use for classification. In C4.5, Gain ration is used which has benefit over Information gain. In ID3, information gain is the decrease in entropy of it produced by splitting the instances based on attribute value [22]. Information gain of node A is calculated using following Equation 1.

$$InformationGain(N,A) = Entropy(N) - \sum_{Values(A)} \frac{|N_i|}{|N|} Entropy(N) \quad (1)$$

Where $A$ is a node, and $N_i$ is subset of N instances for which attribute A has value i. N is set of instances at that node.
Entropy of the set N is computed by using following Equation 2

$$Entropy = - \sum_{i=1}^{No.of classes} \pi log_2 \pi \quad (2)$$

Where, $\pi$ is proportion of instances in N that have their ith class value as output feature. The node of highest information is selected for further process. C4.5 deal with continuous attribute values, missing values and prevent over fitting. For continuous feature value, two branches are created based on threshold rate which finest splits them into two. In C4.5 algorithm, the missing values replaced by values that appear most common in training dataset.

### 3.2. Random Forest

Ensemble algorithms combine more than one same or different kind of algorithms. Leo Breiman has proposed ensemble of trees is named as Random Forest. He have built predictor ensemble using Random forests. Base classifiers of ensemble are a set of decision trees that rise in randomly selected subspaces of data. Each tree is built using an insertion of randomness; this process is called random forest [23]. In this ensemble, the rate of convergence is not depends on number of noise variables. It adapts sparsity and depends only on the number of robust features. Each tree in ensemble is raised according to random parameter. This feature of ensemble of tree can enable user to achieve extensive gains in classification and regression accuracy. Using this ensemble method, final prediction is obtained by combining over the ensemble. Random forests are competitors to SVM and boosting. Random forests are accurate, fast and can handle large number of samples without over fitting [24]. Initially, at each node every tree in the set is formed by first choosing random attribute for splitting purpose by calculating the best split criterion. The CART methodology which maximizes size without pruning is used to grow trees [25]. This randomization of subspace scheme is combined with bagging for re-sampling the training dataset. In bagging, on each sampling with replacement of dataset are used to grow a new individual tree. This predictor using Random forest consists of many randomized base regression trees. These regression trees are denoted by rn(X, $\Phi_m$, Dn), m $\geq$ 1, in which $\Phi_1, \Phi_2, \ldots$ are outputs of a randomizing variable $\Phi$. These trees are combined to form the aggregated regression estimate using Equation 3

$$r_n(X, D_n) = E_\Phi[r_n(X, \Phi, D_n)] \quad (3)$$

Where, $E_\Phi$ denotes expectation with respect to $\Phi$ random parameter, provisionally on $X$ and $D_n$ dataset. The expectation $E_\Phi$ is assessed by Monte Carlo which generates $M$ random trees, and calculating the average of the individual outcomes. The variable $\Phi$ is random which define how the sequential cuts are performed.While constructing the single trees, $\Phi$ is used select the position of splitting coordinator.

# 4. Proposed Ensemble Classifier

In this section, the proposed ensemble classifier for intrusion detection system have explained in detail. Two very known decision trees have combined to form ensemble classifier which offers very high accuracy with less false negatives. For reducing training time, relvant fetaure selection is essential. Genetic search algorithm have used for selection of relevant features. In below subsections, ensemble classifier and feature selection procedure have explained in detail.

## 4.1. Selection of Relevant Features

In this paper, we have used the NSL KDD training dataset to train base classifiers and ensemble classifier. In original training dataset, there are 41 attribute of packets over network. Most of the attribute are derived from other attributes and most of the features are not relevant. These non-relevant attribute take more time for building model. These irrelevant attribute should be removed to reduce model building time. For this purpose, there is need for preprocessing of dataset. Many methods are available for attribute selection, but manual method is usually chosen. In manual method, human understaning of dataset and online packets over internet. In this case, it is very difficult to select relevant attribute by human method. For selection of relevant attribute there is need of optimization techniques to select important features. Genetic algorithm selects the most informative attributes from training and testing dataset. Genetic Search optimization algorithm have used to select relevant attributes. The fifteen relevant and most important features have listed in Table 1.

Table 1: Relevant attributes using Genetic search algorithm

| Sr.No. | Relevant Features | Sr. No. | Relevant Features |
|---|---|---|---|
| 1 | Flag | 9 | Src_serro_rat |
| 2 | Src_bytes | 10 | Same_srv_rate |
| 3 | Dst_bytes | 11 | Diff_srv_rate |
| 4 | Wrong_fragmant | 12 | Dst_host_count |
| 5 | Hot | 13 | Dst_host_srv_diff_host_rate |
| 6 | Logging_in | 14 | Dst_host_serror_rate |
| 7 | Num_file_creation | 15 | Dst_host_srv_serror_rate |
| 8 | Count | | |

## 4.2. Proposed Ensemble classifier

Ensemble classiifer is built using differrent base classifiers. Bagging and Boosting are examble of ensemble classifier in which base classifiers are trained on different training datasets. In combination method, differrent base classifiers can be combined. We can combine different base classifiers using differnet combination rules. In voting combination method, different autonomous base classifiers from same or different family of classifiers can be combined to yield benefit of each base classified. For precise classification, selection of base classifier is very important task. In this paper, we have proposed ensemble classifier using two decision trees. C4.5 and Random forest trees have been selected and implemented as a base classifiers. These both decision trees are mostly used in decision making.

These decision trees have their own advantages and disadvantages. For taking benifits of these trees, they have been selected as base classifiers. The working principles of both decision trees have explained in section 4.2 in detail.

We have used JAVA language for implementation of intrusion detection system. For feature selection, WEKA machine learning tool have used. For combination of base classifier, we taken support of WEKA in built classes.In first phase, C4.5 haveused for intrusion detection system. The size of tree was 320 with 174 leaves and it took 11.11 seconds for model building and 0.33 seconds for testing on training dataset. In second phase, Random forest decision tree had applied for intrusion detection system. It took 80.47 seconds for model building time. The classification accuracies of individual base classifiers and combined ensemble classifier have given in 2. According to Table 2 and Figure 2, it can be observed that C4.5 decision tree as base classifier offers better classification accuracy than Random Base classifier on test dataset. Random forest decision tree as base classifier offers fine classification accuracy on training dataset and cross validation.

Finally, these two base classifiers are combined using four rules of combination. These two base classifiers are combined using different methods in which outputs are combined using combination rules. The overall architecture of proposed intrusion detection system is depicted by Figure 1.It is found that Sum rule is strongest non-trainable combination types which is also called as average rule. In this method, an average of confidence score of a certain class through individual base classifier is calculated to obtain the final score of the class. Equation 4 is used to obtain final score.

Table 2: Classification accuracy of individual base classifiers

| Classifiers | Training in % | C.V. in % | Testing in % |
|---|---|---|---|
| C4.5 | 99.7984 | 99.6991 | 79.0809 |
| Random Forest | 99.9103 | 99.7174 | 78.4155 |
| Ensemble (Average of probability) | 99.8111 | 99.6872 | 79.7862 |
| Ensemble (Product of probability) | 99.8111 | 99.6864 | 79.7862 |
| Ensemble (Minimum probability) | 99.8111 | 99.6864 | 79.7862 |
| Ensemble (Maximum probability) | 99.8111 | 99.6872 | 79.7862 |

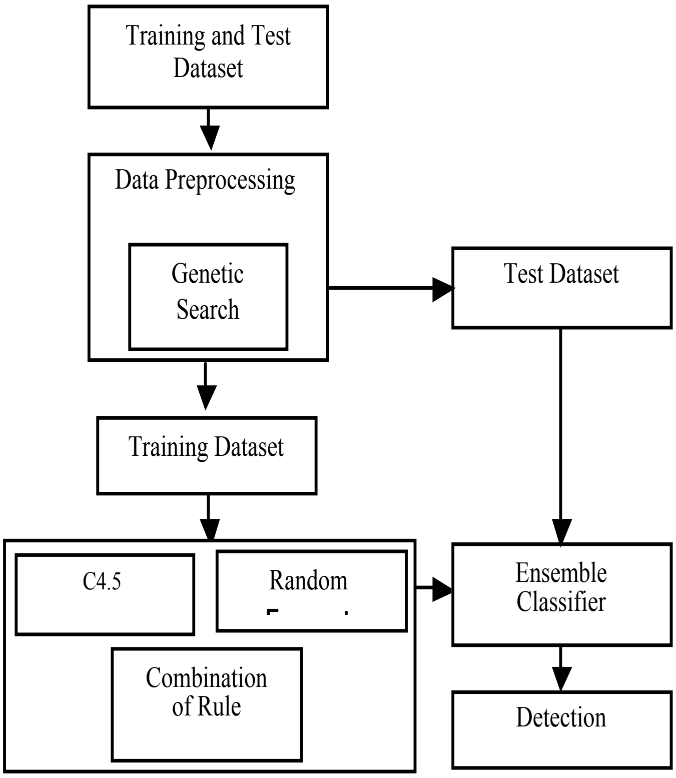

Figure 1: Intrusion Detection System Architechture

$$r_n = \frac{1}{M} \sum_{m=1}^{M} P_m^n \tag{4}$$

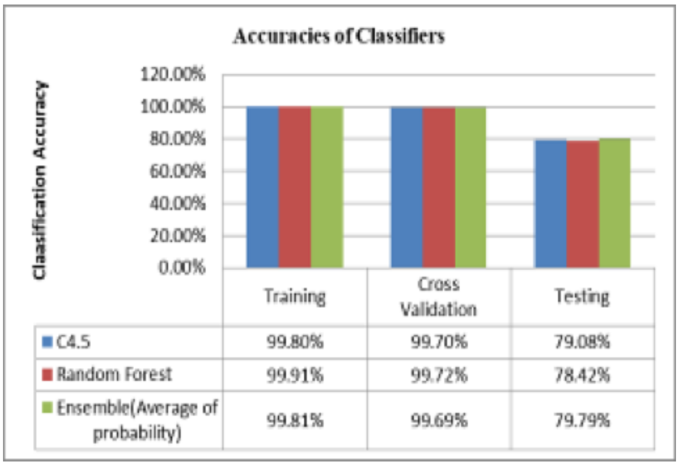

Figure 2: Classification accuracy of Classifiers

Where $P_m^n$ is the posterior score of class $n$ gained from classifier $m$ for any data instances. Where $r_n$, be the combined score of class $n$.

Equation 5 is used to combine classifier using product rule of combination method which is very delicate to the outliers present in the ensemble.

$$r_n = \frac{1}{M}\Pi_m^M P_m^n \tag{5}$$

Equation 6 is used to combine classifier using minimum rule. In this method, final score is computed as the minimum of participating scores.

$$r_n = MINmP_mn \tag{6}$$

Equation 7 is used to combine base classifiers using Maximum combination rule. It just takes the maximum between the classifiers output scores.

$$r_n = MAXmP_mn \tag{7}$$

To combine two classifers, 84.74 seconds taken for model building.

## 5. Results and Discussions

Experiments have conducted on Intel (R) CORE$^{TM}$ i5-3210M CPU with 8 GB RAM, 2.50GHz and 64-bit Operating system. Individual classifiers have implemented and assessed in term of true and fasle positives and classification accuracy. The classification accuracies of ensemble classifier using different combination rules are given in Table 2. According to Table 2 and Figure 2, it can be observed that ensemble classifier with all four combination rules offers better accuracy as compared to both base classifiers on test dataset. On training dataset, ensemble classifier offers less classification accuracy as compared to Random forest on training dataset. On cross validation, ensemble classifier offers better classification accuracy than C4.5 base classifier.

The True positives and false positive of base and ensemble classifiers are listed in Table 3. According to Table 3, ensemble classifier offer very good true positive rate on testing dataset. It is also can be observed that Random Forest Tree gives lowest false positives rate on training dataset. Confusion matrix of each Individual classifier on training dataset is tabulated in Table 4. According Table 4, Random Forest is better classifiers on training dataset. On the other hand, ensemble classifier is best on testing dataset.

## 6. Conclusions

For preventing attacks on computers in network, there is a need of strong security mechanism. The IDS is powerful tool to detect internal intruders who illegally access to the computers of colleagues. It detects malicious activities by analyzing the packets to prevent damage from the attack. In literature, number of attempts has been made to implement intrusion detection systems. The decision trees are predictive models which apply in statistics, data mining and machine learning. In data mining, the decision trees are very popular and widely used for several problems. In this paper, a novel IDS have

Table 3: Confusion matrices

| Classifiers | C4.5 | | Random Forest | | Ensemble Classifier | |
|---|---|---|---|---|---|---|
| | Normal | Anomaly | Normal | Anomaly | Normal | Anomaly |
| Normal | 67210 | 133 | 67276 | 67 | 67222 | 121 |
| Anomaly | 121 | 58509 | 46 | 58584 | 117 | 58513 |

Table 4: Classification accuracy of individual base classifiers

| Classifier | True positive on Training | False Positive on Training | True positive on CV | False Positive on CV | True positive on Testing | False Positive on Testing |
|---|---|---|---|---|---|---|
| J48 | 0.998 | 0.002 | 0.997 | 0.003 | 0.791 | 0.165 |
| Random Forest | 0.999 | 0.001 | 0.997 | 0.003 | 0.784 | 0.170 |
| Ensemble (Average of Probability) | 0.998 | 0.002 | 0.997 | 0.003 | 0.798 | 0.160 |

proposed using ensemble of two classiifer. C4.5 decision tree and Random Forest have selected as a base classifiers of ensemble. The proposed intrusion detection system is formulated by combining the advantages of both C4.5 and Random Forest decision trees. The NSL KDD training dataset have used for training the base and ensemble classifiers. The irrelevant attribute should be removed for reducing model building time. For this purpose, there is need of pre-processing of dataset. Genetic Search optimization algorithm have used for selection of relevant attributes. The performance of proposed system has evaluated in terms of classification accuracy, true positives and false positives. The experimental results show that ensemble classifier with all four combination rules offers better accuracy as compared to both base classifiers on test dataset. On training dataset, ensemble classifier offers less classification accuracy as compared to Random forest on training dataset. On cross validation, ensemble classifier offers better classification accuracy than C4.5 base classifier. It is also observed that ensemble classifier offer very good true positive rate on testing dataset. It is also can observe that Random Forest Tree gives lowest false posistive rate on training dataset.

## References

[1] E. Bauer and R. Kohavi, “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,” *Machine learning*, vol. 36, no. 1-2, pp. 105–139, 1999.

[2] T. G. Dietterich, “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization,” *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.

[3] H. Yao, D. Fu, P. Zhang, M. Li, and Y. Liu, “Msml: A novel multilevel semi-supervised machine learning framework for intrusion detection system,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1949–1959, 2018.

[4] Y. Xiao, C. Xing, T. Zhang, and Z. Zhao, “An intrusion detection model based on feature reduction and convolutional neural networks,” *IEEE Access*, vol. 7, pp. 42 210–42 219, 2019.

[5] W. Alhakami, A. ALharbi, S. Bourouis, R. Alroobaea, and N. Bouguila, “Network anomaly intrusion detection using a nonparametric bayesian approach and feature selection,” *IEEE Access*, vol. 7, pp. 52 181–52 190, 2019.

[6] J. Jingping, C. Kehua, C. Jia, Z. Dengwen, and M. Wei, “Detection and recognition of atomic evasions against network intrusion detection/prevention systems,” *IEEE Access*, vol. 7, pp. 87 816–87 826, 2019.

[7] P. Wei, Y. Li, Z. Zhang, T. Hu, Z. Li, and D. Liu, "An optimization method for intrusion detection classification model based on deep belief network," *IEEE Access*, vol. 7, pp. 87 593–87 605, 2019.

[8] N. Sengupta, J. Sen, J. Sil, and M. Saha, "Designing of on line intrusion detection system using rough set theory and q-learning algorithm," *Neurocomputing*, vol. 111, pp. 161–168, 2013.

[9] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1690–1700, 2014.

[10] W. Feng, Q. Zhang, G. Hu, and J. X. Huang, "Mining network data for intrusion detection through combining svms with ant colony networks," *Future Generation Computer Systems*, vol. 37, pp. 127–140, 2014.

[11] F. Kuang, W. Xu, and S. Zhang, "A novel hybrid kpca and svm with ga model for intrusion detection," *Applied Soft Computing*, vol. 18, pp. 178–184, 2014.

[12] H. E. Ibrahim, S. M. Badr, and M. A. Shaheen, "Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems," *arXiv preprint arXiv:1210.7650*, 2012.

[13] S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," in *The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) l*. IEEE, 2005, pp. 250–257.

[14] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguousand discontiguous system call patterns," *IEEE Transactions on Computers*, vol. 63, no. 4, pp. 807–819, 2013.

[15] M. N. Mohammad, N. Sulaiman, and O. A. Muhsin, "A novel intrusion detection system by using intelligent data mining in weka environment," *Procedia Computer Science*, vol. 3, pp. 1237–1242, 2011.

[16] J. Hu, X. Yu, D. Qiu, and H.-H. Chen, "A simple and efficient hidden markov model scheme for host-based anomaly intrusion detection," *IEEE network*, vol. 23, no. 1, pp. 42–47, 2009.

[17] M. Panda, A. Abraham, and M. R. Patra, "A hybrid intelligent approach for network intrusion detection," *Procedia Engineering*, vol. 30, pp. 1–9, 2012.

[18] C.-J. Chung, P. Khatkar, T. Xing, J. Lee, and D. Huang, "Nice: Network intrusion detection and countermeasure selection in virtual network systems," *IEEE transactions on dependable and secure computing*, vol. 10, no. 4, pp. 198–211, 2013.

[19] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[20] B. Lakshmi, T. Indumathi, and N. Ravi, "A study on c. 5 decision tree classification algorithm for risk predictions during pregnancy," *Procedia Technology*, vol. 24, pp. 1542–1549, 2016.

[21] R. Quinlan and R. Kohavi, "Decision tree discovery." Data Mining, 1999.

[22] U. Bashir and M. Chachoo, "Performance evaluation of j48 and bayes algorithms for intrusion detection system," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 9, no. 4, 2017.

[23] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, no. Apr, pp. 1063–1095, 2012.

[24] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.