RANDOM FOREST, SUPPORT VECTOR MACHINE AND NEAREST CENTROID METHODS FOR CLASSIFYING NETWORK INTRUSION

Sanjiban Sekhar Roy[1], Dishant Mittal [1], Marenglen Biba [2] ,Ajith Abraham[3]

[1] **School of Computer Science and Engineering, VIT University,Vellore, India**
[2] **Department of Computer Science ,University of NewYork,Tirana,Albania**
[3] **Machine Intelligence Research Labs (MIR Labs), Washington, 98071-2259, USA**

Corresponding author: Sanjiban Sekhar Roy, sanjibanroy09@gmail.com

ABSTRACT: Software systems that are capable of controlling a network of computers from malicious intervention are known as intrusion detection systems (IDS).Constantly changing and the complicated nature of intrusion activities on computer networks cannot be dealt with IDSs that are currently operational. This paper proposes a Random Forests method based on the averaging technique for detection of different intrusion types. This proposed method works as a predictive model to detect the types of intrusion attacks. The experimental results indicate that the intended model compared with decision trees, support vector classification models and nearest centroid classification. Proposed model outperforms support vector classification model and nearest centroid classification model.Several comparison metrics have been used like accuracy, the detection rate and false alarm for performance analysis.
KEYWORDS: Intrusion detection system, Random forests, support vector classification, Accuracy, Detection rate, False Alarm.

1. INTRODUCTION

An intrusion detection system (IDS) supervises the network system for suspicious activities and if it detects one, it issues a report to the network management station. In the last few years there has been a step increment in the number of intrusion attacks and as years passed, it has become equally important the task of information assurance by building appropriate intrusion detection systems. The threats such as spyware, hacking, malicious software (like worms and Trojans) is remarkable defiance in regard to maintaining the security of computers and networks. To alleviate this problem, IDS have been drastically deployed into a network environment to identify intrusions.

There are mainly two types of IDS, network based (NIDS) and host based (HIDS). A NIDS acts by placing it at critical points within the network so that it can supervise the signals arriving and leaving from each device in the network. Once the attack is confirmed a notification is sent to the administrator. Host intrusion detection system works on each of the hosts on the network separately. A HIDS supervises the incoming and outgoing traffic from that particular device only and issues a notification if a suspicious activity is detected.

Machine learning techniques are heavily being adapted and developed in intrusion detection to enhance the efficacy of the systems [RV16] and in other applications as well [R+15]. Suthaharan [Sut12] in his work stated that due to the large size and redundant data in the datasets the computation cost of the machine learning methods increases drastically. They proposed ellipsoid-based technique which detects anomalies and side by side cleans the dataset. The research of Chandrasekhar and Raghuveer [CR13] deals with intrusion detection technique which is a combination of k means clustering, neuro-fuzzy and radial basis support vector machine. In their technique, firstly k-means clustering is used to spawn the training subsets, on them various neuro fuzzy models are trained, after that a vector used by svm classification is generated and finally classification task is carried by radial SVM technique.

We propose a method that is based on the classification algorithm named as random forests and use it to detect the intrusions. Random forest is based on ensemble approach and is closely related to decision trees and nearest neighbour methods that are widely used in the task of intrusion detection. Random forest initiates with decision tree, which can be said to be a weak learner approach. A random forest creates a strong learner by combining trees which were stated as weak learners. Random forest works better than decision trees when the number of samples is more [A+15]. In random forests features are selected arbitrarily after each split, this ensures a higher classification power and greater efficiency. Moreover, this method overcomes the problem of over fitting and also it not only pertains the qualities possessed by decision trees, but by utilizing its paging mechanism and voting scheme it produces better results than decision trees mostly [A+12]. In this paper, we present a model that we implemented an intrusion detection system for classification of intrusion types which outperforms the support vector machine

method and the nearest centroid classification method in terms of accuracy, the detection rate and false alarm. An analysis has been performed for each type of attack mentioned in the dataset that has been utilized for this study.

1.1KDD CUP DATA SET

This dataset is developed from the DARPA packet traces and it comprises of a huge collection of intrusions that take place in a military network environment which can be downloaded from the site (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html). The original KDD dataset is comprised of 4898431 instances, each of which contains 41 features and is classified with exactly one particular attack type. The type can be either normal or attack. Every instance that is not labelled as normal is considered to be attack. KDD CUP 99 has been dramatically used in network attacks. Every attack is comprised in one of the following categories [TK12]:

A)Denial of Service Attack (DOS)
Denial of Service Attack (DOS): Under this category the attacker makes any computing resource or memory resources too much busy so that it becomes incapable of handling any genuine requests. Hence, the machine denies any legitimate users access to machines.

B)Users to Root Attack (U2R)
Users to Root Attack (U2R): Under this category the attacker firstly finds an access to the account of a normal user on the system and further is able to exploit some vulnerable factor so as to obtain the root access to the system.

C)Remote to Local Attack (R2L)
Under this category the attacker not having an account on a machine in the network delivers packets to that machine and exploits some vulnerable factor so that to have an authority as a user of that machine.

D)Probing Attack (PROBE)
Under this category the attacker tries to collect the information related to the computer in the network with an intention of dodging the established security.

Table 1. Attack categories and type

S.No.	Probe	Dos	U2R	R2L
1	Ipsweep	Back	Buffer overflow	ftp_write
2	Nmap	Land	Loadmodule	Guess_passwd
3	Portsweep	Neptune	Perl	Imap
4	Satan	Pod	Rootkit	Multihop
5	-	Smurf	-	Phf
6	-	Teardrop	-	Spy
7	-	-	-	Warezclient
8	-	-	-	Warezmaster

The main aim of detecting the network intrusion are detecting the attacks that occur rarely like U2R and R2L and increase the accuracy, detection rate thereby decreasing rate of false alarm.
The protocols that are used in KDD dataset are UDP, TCP and ICMP.The experiment performed utilizes several machine learning techniques to analyze efficiency and study the pattern of different types of attacks in relation to these protocols.

2. RELATED WORK

Sangakatsanee et al. [SWC11] proposed a supervised machine learning approach and utilized that technique in an intrusion detection system. They conducted experiments for detecting the attack type using various techniques and concluded that the decision tree technique was ahead of all. Using a feature selection criteria they discovered 12 important features that were critically associated with detecting network attacks. Shon et al. [SM07] proposed a slight modified SVM method called as enhanced SVM which incorporates the one class SVM approach and the generic approach so that they could deliver an unsupervised learning with low alarm rate in the same way as supervised SVM approach. Toosi and Kahani have [TK12] integrated several soft computing methods for constructing an intrusion detection system to classify the attacks into various categories. They utilized a combination of neuro fuzzy classifiers for the task of classifying the attacks and genetic algorithm was utilized for optimization of the configuration of the fuzzy decision engine. Helmer et al. [H+00] proposed a distributed intrusion detection system, which was integrated with static and portable agents and some agents like data warehouse. These agents ensured creation, supervision and investigation of spatio-temporal and generic criteria for huge distributed systems. Chen at al. [C+07] implemented svm and artificial neural networks (ANN) with encoding techniques like tf x idf and simple frequency-based scheme for the task of intrusion detection. They concluded that svm with tf x idf scheme performed the best whereas ANN with simple frequency-based scheme performed worst. They utilized KDD 99 dataset for experiments. Horng et al. [H+11] proposed a hybrid approach towards intrusion detection, which combined feature selection, hierarchical clustering and svm technique. They implemented it on KDD cup 99 dataset. The results included improved performance, reduced training time and improved accuracy. Elhag at al. [E15]has utilized fuzzy systems and pairwise learning to raise the detection rate of the intrusions. Li et al. [L+12] proposed a hybrid approach combining clustering, ant colony algorithm and svm for detecting network attacks and classify them into categories. Feature

removal was also conducted and the result was an accuracy of over 98.6 %. Levent Koc al. [KMS12] claimed that a technique known as Hidden Naïve Bayes (HNB) model proves to be efficient in circumstances related to highly correlated feature vectors and the problems which suffer from dimensionality. Their results indicated that HNB performs better than typical Naive Bayes models and models like svm when applied to KDD cup 99 datasets. In their research Kim et al. [KLK14] build a highly accurate misuse detection model related to decision tree and then decomposed the generic training data into small chunks for which one–class svm models were created. Tian et al. [GTX09] presented a method exclusively for host based intrusion systems. The sequence of the shell commands is used to create multiple sequence libraries signifying the normal profile. In detection stage, mining is performed in the pre-recorded data using a sequence matching algorithm and the mined patterns are evaluated for the similarities with respect to historical profile. Finally the smoothed similarities are used to decide on normal or attack conditions. The accuracy was reported high with significantly less detection time. A better accuracy was perceived as compared to existing techniques. KDD 99 dataset was used for experimentation. The reviewed techniques included neural networks, soft computing, swarm intelligence, fuzzy logic. Mukherjee and Sharma [MS12] suggested a feature reduction technique based on feature importance which is used to recognize vital reduced features. Naïve Bayes was applied on the resulting dataset for anomaly detection and they found that with reduced features it was more efficient. Zhang and Zulkernine [ZZ06] proposed a hybrid approach which combines anomaly detection technique with misuse detection and applied it over KDD 99 dataset. They concluded that this hybrid approach performed better than if either anomaly or misuse detection technique is used in isolation.The resultant training and testing times were significantly reduced and descent accuracy was reported. Moskovitch et al. [M+07] presented a technique based on machine learning for categorizing the codes in host based intrusion detection system as malicious or benign. The algorithm is trained by previously detected malicious code. The accuracy was detected to be above 90 %. Das et al. [DSB09] implemented a web intrusion detection system. An unsupervised technique was used for clustering of the queries which helped in detection of intrusions. They used a two phase approach in which the first was related to matching mechanism and in second phase the frequency of the arriving packets was an important criteria for detection. Both phases worked simultaneously. Gao et al. [GTX09] combined genetic algorithm with svm for the task of intrusion detection. This combination ensured the optimization of svm parameters. The results of their implementation specified that the method was effective in detecting the attacks. Through their study they claimed that after including a certain number of classifiers, the accuracy of the system remains intact. Roy et al. also have proposed Intrusion detection systems based on rough set and dominance based rough set approaches respectively [R+12], [R+13], [DAP+14]. They arrived at a conclusion that practical context is a very crucial factor in evaluating the performance of machine learning algorithms. Muda et al. [M+11] proposed a hybrid method which combines k-means algorithm with naïve bayes. The mentioned approach clustered the instances to their corresponding classes. The dataset utilized was KDD 99 dataset. The results indicated a descent accuracy level.

3. CLASSIFICATION MODEL

In general, the category of problems which contains data as well as the additional attributes that we want to predict comes under supervised learning approach. Under supervised learning approach the classification problem comes into account when the instances belong to two or more classes and our intention is to forecast the class of the unlabeled instances. Under the category of supervised learning methods, a technique known as Support vector machines (SVM) holds its place for classification. This method is effective for high dimensional spaces, is memory efficient since it utilizes subset of training data points in the decision function called as support vectors, also it is adroit as for the decision function various kinds of kernel functions can be stated . If the count of features is bigger than the count of samples this technique is liable to give mediocre performance.

A. Suggested Classification Model

The suggested classification model is comprised of random forest algorithm; it advances by building an aggregation of decision trees during the process of training and comes under the category of ensemble methods. This method is based on creating several models in isolation and after that take a mean of the predictions done by each of the models. The output class is computed by taking the mode of the labels calculated by each tree.

B. Mathematical Formulation

Given a set of training data points,

$$D = \{Xi, Yj\}$$

where i =1…n. A weighted neighbourhood concept is applied as shown, given a forest of M trees, prediction made by the m-th tree for X can be written as

$$Tm(X) = \sum_{i=1}^{n} Wim(X)Yi$$

where Wim(X) is equal to 1/km , if X and Xi are in the same leaf in the m-th tree and 0 otherwise, km is the number of training data which fall in the same leaf as X in the m-th tree. The prediction of the entire forest is given by,

$$F(X) = \frac{1}{M}\sum_{m=1}^{M} Tm(X) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{n} Wim(X)Yi$$
$$= \sum_{i=1}^{n}\left(\frac{1}{M}\sum_{m=1}^{M} Wim(X)\right)Yi$$

which signifies that the prediction made by the random forest algorithm is the weighted average of the Yi's, with weights,

$$Wi(X) = \frac{1}{M}\sum_{m=1}^{M} Wim(X)$$

The data points Xi which are contained in the corresponding leaf as X in at least one tree of the forest are termed as the neighbours of X in this analysis.

4. EXPERIMENTAL ANALYSIS

A. Data Preparation

In our experiments, two per cent of original KDD CUP 99 dataset that is approximately 97968 single connection vectors has been utilized. Since KDD CUP 99 dataset contains a vast amount of instances, it is very computationally expensive and requires powerful machines to perform the task. Therefore, a subset of 2% of this dataset is considered for the experimentation. Since the first 97968 original dataset was comprised of only normal, buffer_overflow, loadmodule, perl, neptune and smurf as the class types, to ensure the uniform distribution of all the classes in the considered portion the dataset was shuffled. Each instance consists of 41 features and is classified with one particular attack type mentioned in table 1. Further description regarding available features can be found in. The training data set was considered as the first 70% of the instances, while the testing data comprised of remaining 30% of the instances. From the dataset considered for implementation the distribution of the training and the testing dataset according to the class type is shown in table 3. We have used python programming environment for all the experiments.

Table 2. Sample distribution of training and testing dataset

Class	Training Dataset		Testing Dataset	
	No. of samples	Sample Percentage (%)	No. of samples	Sample Percentage (%)
Normal	13484	13.763	5744	5.863
Probe	569	0.580	261	0.266
Dos	54351	55.478	23301	23.784
U2R	6	0.006	4	0.004
R2L	167	0.170	80	0.081
Total	68577	100	29390	100

B. Algorithm

```
Random Forest Classification ( )
1.data ←read data set
2.data ←convert_features_text_to_integral (data)
3.(train_features,train_intrusion_categories)←
training_function()
4.(test_features,test_intrusion_categorie)←
testing_function()
5.model←randomforests_train(train_features,
train_intrusion_categories)
6.classification_result← randomforests_classify
(train_features)
7.Accuracy ← (TP+TN) / (TP+TN+FP+FN)
8.Detection_Rate ← TP / (TP+FP)
9.False_Alarm ← FP / (FP+TN)
```

5. RESULT AND DISCUSSION

The performance of all the classifiers was computed by utilizing a matrix known as confusion matrix. It is a standard metric for benchmarking the effectiveness and robustness of a classification algorithm. Using the confusion matrix, measures like accuracy, detection rate and false alarm rate have been computed which are the generic criteria for evaluating the performance of the IDS. These metrics have been utilized in a number of studies and they ensure a viable means of deciding the efficiency of the model for detecting the intrusions within systems. For a decent level of performance, the intrusion detection system (IDS) needs high accuracy and detection rate and conversely false alarm rate should be low. These terms are given by the following formulae:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Detection\_Rate = \frac{TP}{(TP + FP)}$$

$$False\_Alarm = \frac{FP}{(FP + TN)}$$

Table 3 represents a matrix known as confusion matrix. True positive (TP) indicates the number of instances having the class label of attack and were correctly classified as an attack. True negative (TN) indicates the number of instances having the class

label of normal and were correctly classified as normal. False positive (FP) indicates the number of instances that have a label of being valid but have been incorrectly classified as intrusion. False negative (FN) indicates the number of instances that were having a label of intrusion but were incorrectly classified as normal by the IDS.

Table 3. Confusion matrix

		Predicted class	
		Class=attack	Class=normal
Actual Class	Class=attack	TP	FN
	Class=normal	FP	TN

A ten-fold cross validation scheme has been followed for computing the confusion matrix corresponding to each kind of attack. This has been accomplished for both training as well as testing data set. Table 4, 5 and 6 signify the best result among the 10 outputs returned after ten-fold cross validation for nearest centroid, svm and random forests, respectively. For simultaneous observation in each table, the confusion matrix for testing and training sets in each type of attack, namely DOS, PROBE, R2L and U2R have been integrated into a single table. The last 3 rows in table 4, 5 and 6 denote accuracy, the detection rate and false alarm rate for the corresponding training or test set. It is apparent that,in our proposed work, Random forest has accomplished dependable peak values for all intrusion types. In case of DOS intrusion, 99.996 % is attained, which is the maximum accuracy, when compared with other techniques.

In case of PROBE intrusion, a very descent accuracy of 99.916 % has been attained. For R2L and U2R, once again the highest accuracy values of 99.914 and 99.947 are reported, respectively. Table 7 represents the experimental results of the three techniques used for intrusion detection over the KDD CUP 99 dataset. The accuracy values were calculated for the testing dataset to elaborate on the efficiency of the compared models. It was concluded that the random forests method outperforms the other two methods.

Table 4. Experimental results obtained for the training and testing dataset using the nearest centroid technique

Class	Types of attacks							
	Dos		Probe		R2L		U2R	
	Train	Test	Train	Test	Train	Test	Train	Test
(TN)	2176	892	2176	892	2176	892	2176	892
(FP)	3344	1365	1504	606	5898	2640	562	241
(TP)	53386	22864	294	132	21	8	1	0
(FN)	0	0	0	0	43	21	2	1
ACCURACY	94.323	94.566	62.154	62.822	26.996	25.273	79.423	78.659
DETECTION RATE	94.105	94.366	16.351	17.886	0.354	0.302	0.177	0.0
FALSE ALARM	60.579	60.478	40.869	40.453	73.049	74.745	20.525	0.0

Table 5.Experimental results obtained for the training and testing dataset using the support vector machine technique

Class	Types of attacks							
	Dos		Probe		R2L		U2R	
	Train	Test	Train	Test	Train	Test	Train	Test
(TN)	12211	5737	12211	5737	12211	5737	12211	5737
(FP)	59	4	766	3	446	0	2	0
(TP)	54000	23112	396	178	72	4	1	0
(FN)	5	189	44	64	94	76	5	4
ACCURACY	99.903	99.335	93.962	98.879	95.788	98.693	99.942	99.930
DETECTION RATE	99.890	99.982	34.079	98.342	13.899	100	33.333	0
FALSE ALARM	0.480	0.069	5.902	0.052	3.523	0	0.0163	0

Table 6. Experimental results obtained for the training and testing dataset using proposed technique

Class	Types of attacks							
	Dos		Probe		R2L		U2R	
	Train	Test	Train	Test	Train	Test	Train	Test
(TN)	13483	5740	13483	5740	13483	5740	13483	5740
(FP)	1	1	0	0	0	3	0	0
(TP)	54351	23299	568	255	167	76	6	1
(FN)	0	0	1	5	0	2	0	3
ACCURACY	99.998	99.996	99.992	99.916	100	99.914	100	99.947
DETECTION RATE	99.998	99.995	100	100	100	98.202	100	100
FALSE ALARM	0.007	0.017	0	0	0	0.052	0	0

Table 7. Accuracy, detection rate and false alarm rate values after a 10-fold cross validation scheme

	Different Methods	Dos		Probe		R2L		U2R	
		Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	NC	94.138	94.423	62.003	62.819	26.368	25.107	79.293	78.003
	SVM	99.386	98.812	93.358	98.230	95.198	92.597	99.022	98.891
	RF	99.726	99.978	99.923	99.901	99.989	99.9101	99.967	99.788
Detection Rate	NC	94.008	94.253	16.164	17.112	0.211	0.109	0.088	0.0
	SVM	99.677	99.762	34.063	98.185	13.793	95.781	33.31	0
	RF	99.947	99.866	99.945	99.921	99.978	98.971	99.901	99.916
False Alarm Rate	NC	60.649	60.757	40.972	40.590	73.129	74.802	20.671	0.1
	SVM	0.483	0.074	5.913	0.061	3.609	1	0.0215	0
	RF	0.007	0.018	0	1	0	0.054	0	0

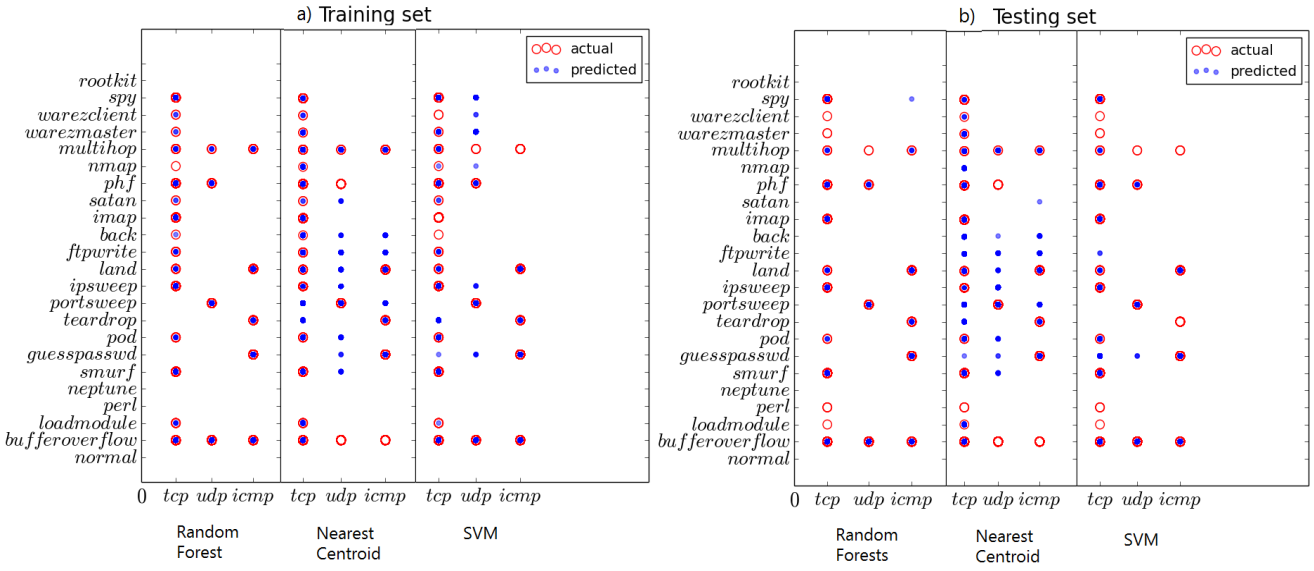

Figure 1. Graphical representations for comparison of the other techniques with our proposed technique (train and test sets)

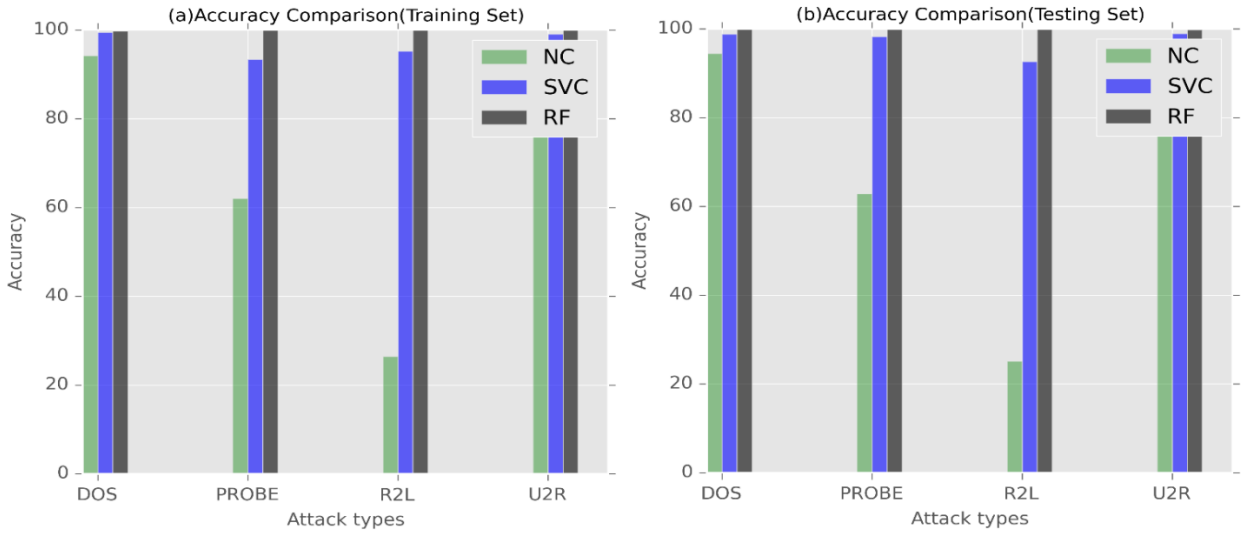

Figure 2. Accuracy comparison of other techniques with our proposed technique (train and test sets)

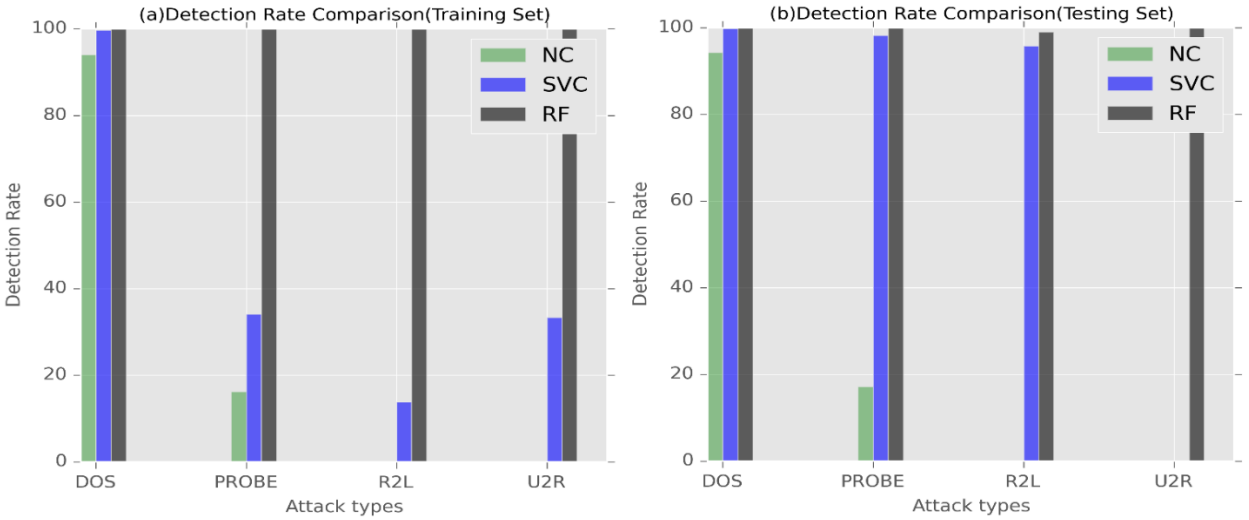

Figure 3. Detection rate comparison of other techniques with our proposed technique (train and test sets)

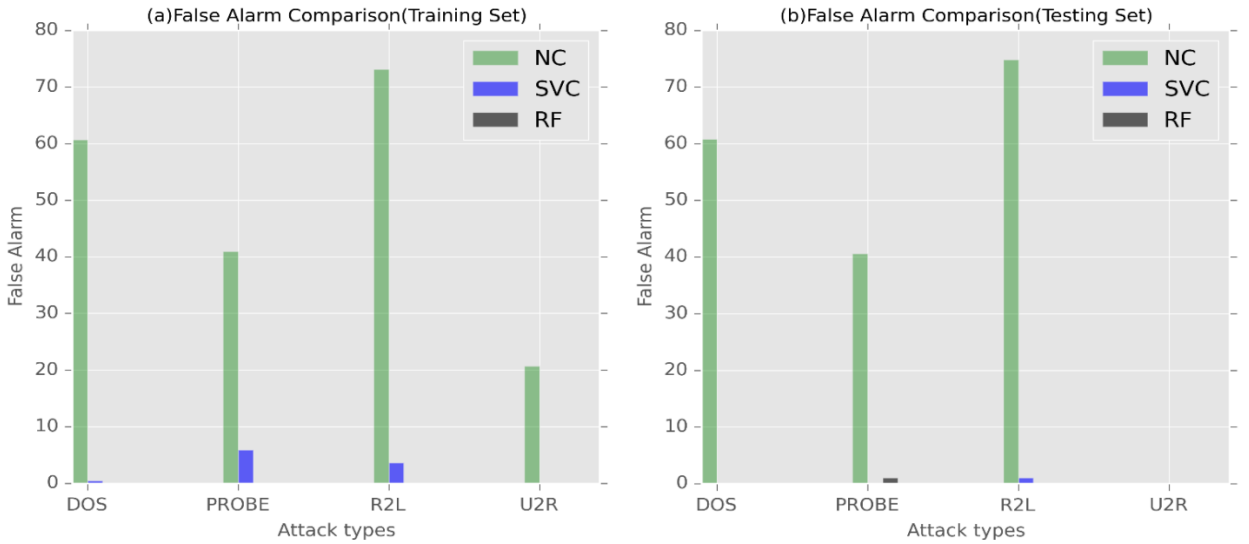

Figure 4. False Alarm comparison of other techniques with our proposed technique (train and test sets)

Fig. 1 depicts a comparison between nearest centroid, svm and random forests. Fig. 1 (a) and (b) have been plotted with respect to the protocol feature. Protocol feature in turn could be any one of three discrete types which are tcp, udp and icmp. For a particular feature the red circles signify actual class labels and blue dots signify predicted class labels. Since red circles signify actual classes and blue dots signify predicted classes, hence, the

more the number of blue points outside the circle, the lesser is the accuracy of that model or the more the number of empty circles the lesser is the accuracy for that model. Fig. 2-4 describe the comparative results of the experimentation in the graphical format for accuracy, detection rate and false alarm rate. In each fig. 2-4, (a) and (b) denote testing and training set results, respectively. These graphs clearly show that the proposed model can provide an efficient intrusion detection tool compared to other equivalent advanced predictive techniques.

CONCLUSION

Intrusion detection is a significant task nowadays for computational methods. In this paper a novel method for detecting intrusion type was introduced. 41 feature vectors were utilized as input to classify attack categories of the KDD Cup dataset successfully. An IDS system that was able to analyse the dynamic and complex nature of intrusion activities has been built. Random forests technique based on the averaging method outperformed the major classification methods support vector machine and nearest centroid. After achieving ten-fold cross validation, the proposed model resulted in the highest accuracy and detection rate values as well as the least alarm rate values. This trend was maintained consistently for PROBE, DOS, U2R and R2L intrusion attacks. The results specify that the classification ability of the proposed model is incomparable to nearest centroid approach and is inherently superior to the support vector machine model. Anomaly detection methods that are based on artificial intelligence are continuously alluring a lot of attention from the research community. Through this paper, we were able to successfully perform a comparison study and come over with a state of the art method in intrusion detection system.

REFERENCES

[A+12] **J. Ali, R. Khan, N. Ahmad, I. Maqsood** - *Random forests and decision trees*, IJCSI International Journal of Computer Science Issues, vol. 9, no. 5, 2012.

[A+15] **S. Adusumilli, D. Bhatt, H. Wang, V. Devabhaktuni, P. Bhattacharya** - *A novel hybrid approach utilizing principal component regression and random forest regression to bridge the eripod of GPS outages*, Neurocomputing, 2015.

[CR13] **A. Chandrasekhar, K. Raghuveer** - *Intrusion detection technique by using k-means, fuzzy neural network and svm classifiers*, In Computer Communication and Informatics (ICCCI), 2013 International Conference, pp. 1-7.

[C+07] **Y. Chen, L. Dai, Y. Li, X. Q. Cheng** - *Building lightweight intrusion detection system based on principal component analysis and C4. 5 algorithm*, In Advanced Communication Technology, The 9th International Conference on IEEE. vol. 3, pp. 2109-2112, 2007.

[DAP14] **T. K. Das, D. P. Acharjya, M. R. Patra** - *Business Intelligence from Online Product Review-A Rough Set Based Rule Induction Approach*, 2014 International Conference on Contemporary computing and Informatics (IC3I-2014), 27-29 Nov, 2014, pp. 800-803, Mysore, India, IEEE Xplore, DOI:10.1109/IC3I.2014.7019662, 2014.

[DSB09] **D. Das, U. Sharma, D. K. Bhattacharyya** - *A Web Intrusion Detection Mechanism based on Feature based Data Clustering*, In Advance Computing Conference, 2009.

[GTX09] **M. Gao, J. Tian, M. Xi** - *Intrusion detection method based on classify support vector machine*, In Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on, vol. 2, pp. 391-394, IEEE, 2009.

[H+00] **G. Helmer, J. S. Wong, V. Honavar, L. Miller** - *Automated discovery of concise predictive rules for intrusion detection*, Journal of Systems and Software, vol. 60, no. 3, pp. 165-175, 2002.

[H+11] **S. J. Horng, M. Y. Su, Y.H. Chen, T. W. Kao, R.J. Chen, J.L. Lai, C. D. Perkasa** - *A novel intrusion detection system based on hierarchical clustering and support vector machines*, Expert systems with Applications, vol. 38, no. 1, pp. 306-313, 2011.

[KLK14] **G. Kim, S. Lee, S. Kim** - *A novel hybrid intrusion detection method integrating anomaly detection with misuse, Expert Systems with Applications* 41.4, 1690-1700, 2014.

[KMS12] **L. Koc, T. A. Mazzuchi, S. Sarkani** - *A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier*, Expert Systems with Applications, vol. 39, no. 18, pp. 13492-13500, 2012.

[L+12] **Y. Li, J. Xi, S. Zhang, J. Yan, X. Ai, K. Dai** - *An efficient intrusion detection system based on pport vector machines and gradually feature removal method*, Expert Systems with Applications, vol. 39, no. 1, pp. 424-430, 2012.

[MS12] **S. Mukherjee, N. Sharma** - *Intrusion detection using naive Bayes classifier with feature reduction*, Procedia Technology, vol. 4, pp. 119-128, 2012.

[M+07] **R. Moskovitch, S. Pluderman, I. Gus, D. Stopel, C. Feher, Y. Parmet, Y. Elovici** - *Host based intrusion detection using machine learning*, In Intelligence and Security Informatics, IEEE, pp. 107-114, 2007.

[M+11] **Z.Muda, W. Yassin, M. N. Sulaiman, N. I. Udzir** - *Intrusion detection based on K-Means clustering and Naïve Bayes classification*, In Information Technology in Asia (CITA 11), 2011 7th International Conference on, pp. 1-6.

[RV16] **S. S. Roy, V. M. Viswanatham** - *Classifying Spam Emails Using Artificial Intelligent Techniques*. In International Journal of Engineering Research in Africa, vol. 22, pp. 152-161. Trans Tech Publications, 2016.

[R+12] **S. S. Roy, O. Jadhav, S. Chakraborty, V. M. Viswanatham** - *Multicriteria Decision Analysis for Intrusion Detection Data*, Proceedings of International Conference on Advances in Computing. Springer India, pp. 667-672, 2012.

[R+13] **S. S. Roy, V. M. Viswanatham, P. V. Krishna, N. Saraf, A. Gupta, R. Mishra** - *Applicability of Rough Set Technique for Data Investigation and Optimization of Intrusion Detection System*, Quality, Reliability, Security and Robustness in Heterogeneous Networks. Springer Berlin Heidelberg, pp. 479-484, 2013.

[R+15] **S. S. Roy, D. Mittal, A. Basu, A. Abraham** - *Stock Market Forecasting Using LASSO Linear Regression Model*. In Afro-European Conference for Industrial Advancement, pp. 371-381. Springer International Publishing, 2015.

[Sut12] **S. Suthaharan** - *An iterative ellipsoid-based anomaly detection technique for intrusion detection systems*, In Southeast on, Proceedings of IEEE, pp. 1-6, 2012.

[SM07] **T. Shon, J. Moon** - *A hybrid machine learning approach to network anomaly detection*, *Information Sciences*, vol. 177, no. 18, 3799-3821, 2007.

[SWC11] **P. Sangkatsanee, N. Wattanapongsakorn, C. Charnsripinyo** - *Practical real-time Intrusion detection using machine learning approaches*, Computer Communications, vol. 34, no. 18, pp. 2227-2235, 2011.

[TK12] **A. N. Toosi, M. Kahani** - *A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers*, Computer communications, vol. 30, no. 10, pp. 2201-2212, 2012.

[ZZ06] **J. Zhang, M. Zulkernine** - *Anomaly based network intrusion detection with nsupervised outlier detection*. In Communications, 2006. ICC'06. IEEE International Conference on (vol. 5, pp. 2388-2393). IEEE, (2006, June).