INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue1)
Available online at: www.ijariit.com

Novel approach for IDS Classification Support vector machine and adaptive boost

Er. Aanchal Kumar
Student
Dept. of C.S.E., Rayat Bahra Group of Institutes (P.T.U), Patiala, India.
amiable13aanchal@gmail.com

Er. Jaspreet Kaur
Assistant Professor
Dept. of C.S.E., Rayat Bahra Group of Institutes (P.T.U), Patiala, India.
jaspreetkaur.rb2@gmail.com

Er. Inderpreet Kaur
Assistant Professor
Dept. of C.S.E., Rayat Bahra Group of Institutes (P.T.U), Patiala, India.
inderpreet.kaur029@gmail.com

Abstract— we perform three sets of experiments. From the first experiment, the systems are trained using all the 41 features. The second experiment where we perform feature selection by using Gain Ratio as to select the best features instead of using all the 41 features and perform the experiment with Linear SVM, SGD and Adaptive Boost and compare the results. The third experiment where we perform feature extraction by using Gain Ratio as to select the best features instead of using all the 41 features and perform the experiment with Linear SVM, SGD and Adaptive Boost and compare the results.

Keywords: Features, SVM, KDD 99, SGD.

I. INTRODUCTION

With the extension of business internationally the use of World Wide Web is also expanded for all kind of business environment and as well as in daily life, the need of safe information is expanded. All the world wide structures are nowadays based upon computer networks. To safeguard the flow of confidential information over the network, network security has become a necessary requirement of modern society. One of the most extraordinary tasks is the detection of Intrusion over the network to safeguard the confidential information by the attackers [1].

Well-organized intrusion detection is required as a defence of the network system to predict the attacks over the network. A characteristic choice and categorization on the basis of Intrusion Detection model is available, by carrying out feature choice, the dimensions of NSL-KDD data set is minimized then by pertaining machine learning perspective, we are able to construct Intrusion detection model to find attacks on structure and enhance the intrusion detection with the use of captured data [1] With the expansion of number of new unseen attacks the ambition of this model is to establish a system for intrusion detection, and the model will be competent of detecting new and prior unseen attacks with the use of prime signatures and the characteristic of known attacks. [2].

Attackers always get attracted to the significant and profitable information and are always responsible to uttermost attacks over the network. Intrusion can get into the structure or system server by sending the malicious packet to the user system and then modifying or altering any confidential information or significant information and sends packet over the network for illegal ambition called attack. Because of the system weakness or vulnerability such as user misuse,

Mis-configuration; intrusion occurs over the system server. An intelligent intrusion can also be made by putting together multiple susceptibility. There are millions of big servers in a global network and has a great amount of on-line services running in the system which attracts more attackers due to which there is need of intelligent intrusion detection model; acts as a defence for their network system.

II. LITERATURE REVIEW

In [3] Carl Livadas et al In this paper main focus is on monitoring network traffic, anomalies detection and cyber-attack traffic patterns, and, a posteriori, cyber-attacks combat and effects mitigation. Contrary to such approaches, they advocate proactively detecting and identifying botnets prior to their being used as part of a cyber-attack. In this paper, machine learning-based classification techniques is based on machine learning is used in present work to identify the command and control (C2) traffic of IRC-based botnets —Internet Relay Chat (IRC) uses collective commands for host compromise. There task is divided into two stages: (I) am distinguishing between IRC and non-IRC traffic, and (II) distinguishing between botnet and real IRC traffic. For Stage I, the performance of J48, naive Bayes, and Bayesian network classifiers is compared and the features are identified that achieve good overall classification accuracy, and training set size, the classified sensitivity is determined. While sensitive to the

training data and to characterize communication flows attributes are used, classifiers based on machine learning show promise in identifying IRC traffic. Trickier is the classifier used in Stage II, since botnet as the accurately labelling IRC traffic and non-botnet is challenging. The labelling flows are currently exploring as suspicious and non-suspicious based on telltales of hosts being compromised.

In [4] Wenke Lee et al. In this paper, their research for intrusion detection system is in developing general and systematic methods. To construct detection model, framework consists of classification, association rules, and frequency episodes programs is used and the effectiveness of classification models is demonstrated by the experiment on send mail system call data and network tcpdump data. An overview on two general data mining algorithms implemented: the association rules algorithm and the frequent episodes algorithm which are used to compute the intra- and inter- audit record patterns describing program or user behaviour. To the challenges of both efficient learning (mining) and real-time detection are met, an agent-based architecture for intrusion detection systems is proposed.

In [5] l data mining techniques can be in intrusion detection in which ADAM(Audit Data Analysis Mining) system is used as a testbed and studies the design and experiences with ADAM. Traditionally, intrusion detection systems based on the attack characterization and the tracking to see whether system activity matches the characterization. Explains how new intrusion detection systems are making their appearances in the field which are based on data mining.

In [6] S. Mukkamala et al. This paper describes its approach for intrusion detection by using neural networks and support vector machines. The main concept is based on the ideas to built a classifier using a set of relevant features for recognizing anomalies and known intrusion in real time and user behavior on the system is described to discover useful pattern or features. DARPA uses a benchmark data set for designing the KDD (knowledge discovery and data mining) and demonstrate that to detect intrusion; efficient and accurate classifiers can be built. Further comparison between the neutral network based intrusion detection system and support vector machine based intrusion detection system performance is done.

In [7] Yi Hu et al. In this paper, in a Database System, for detecting malicious transactions the proposed approach used is data mining approach and concentrates on this approach dependencies on data items. From the database log, miner data dependency for mining data correlations is designed. The malicious transactions is identified when the the data dependencies mined are not compliant by the transactions. The proposed method illustrated by the experiment works effectively for detecting malicious transactions in the database provided certain data dependencies exist.

In [8] Paul Dokas et al. In this paper an overview for identifying known intrusions in building rare class prediction models and anomaly/outlier detection schemes and variation for detecting novel attacks having unknown nature. Experimental results shows that rare class predictive models on the KDDCup'99 dataset have much more efficiency in the detection of intrusive behaviour and DARPA 1998 data set experimental result shows that the new techniques have great promise in detecting novel intrusions.

In [9] Andrew Honig et al. In this paper, Adaptive Model Generation (AGM) is used as real time architecture for implementing data mining-based intrusion detection systems. The associated problem with data mining based-IDS are solved using this architecture by automating the collection of data, the generation and deployment of detection models and the real-time evaluation of data. Specific examples of system components including auditing sub-systems, model generators for misuse detection and anomaly detection and support for virtualization and correlation of multiple audit sources.

In [10] J.E. Dickerson et al. Describes an anomaly- based intrusion detection system that is Fuzzy Intrusion Recognition Engine (FIRE) which uses fuzzy logic to assess whether on a network, malicious activity is taking place. The network input data and help expose metrics is processed by using simple data mining techniques for anomaly detection significance and fuzzy sets are than evaluated by using fuzzy analysis engine in the FIRE for security administrations. In this paper, FIRE architectures and its role are described.

III. OBJECTIVES

To reduce the dimension of feature extraction by KPCA (Kernel based principle component analysis) and correlation feature selection method. To improve the learning model by SVM RBF Kernel and to compare the existing method by our method.

IV. PROPOSED METHDOLOGY

The methodology of designing the proposed scheme is divided into three phases: Feature Selection, Feature Extraction and Normal

3.4.1 Feature Selection: In the feature selection phase the following steps takes place as shown in fig

Step1: KDD-99 dataset with 41 features
Step2: Feature Selection by information gain & correlation method.
Step3: Input the feature & labels into SVC, SGD & adaptive boost &make three models.
Step4: Perform the test on these models & calculate the precision, recall, and accuracy.

3.4.2 Feature Extraction: In this feature extraction phase, the following steps take place as shown in the fig.

Step1: KDD-99 dataset with 41 features
Step2: Feature extraction with KPCA
Step3: Input the feature & labels into SVC, SGD & adaptive boost & make three models.
Step4: Perform the test on these models and calculate the precision, recall & accuracy.

3.4.3 Normal: In this phase the following steps takes place as shown in fig 4.

Step1: KDD-99 dataset with 41 features

Step2: Input the feature & labels into SVC, SGD & adaptive boost &make three models.
Step3: Perform the test on these models and calculate the precision, recall accuracy.

3.5 Performance Metrics.
The List below shows important performance parameters chosen to analyse the results.
1. Precision
2. Recall
3. Accuracy
Precision: It is the number of instances correctly classified as class X among those classified as class X. Simply put, the precision address the following questions. Based on prediction, how likely is it that the prediction be true?

$$\text{Precision} = \frac{TP}{TP+FN}$$

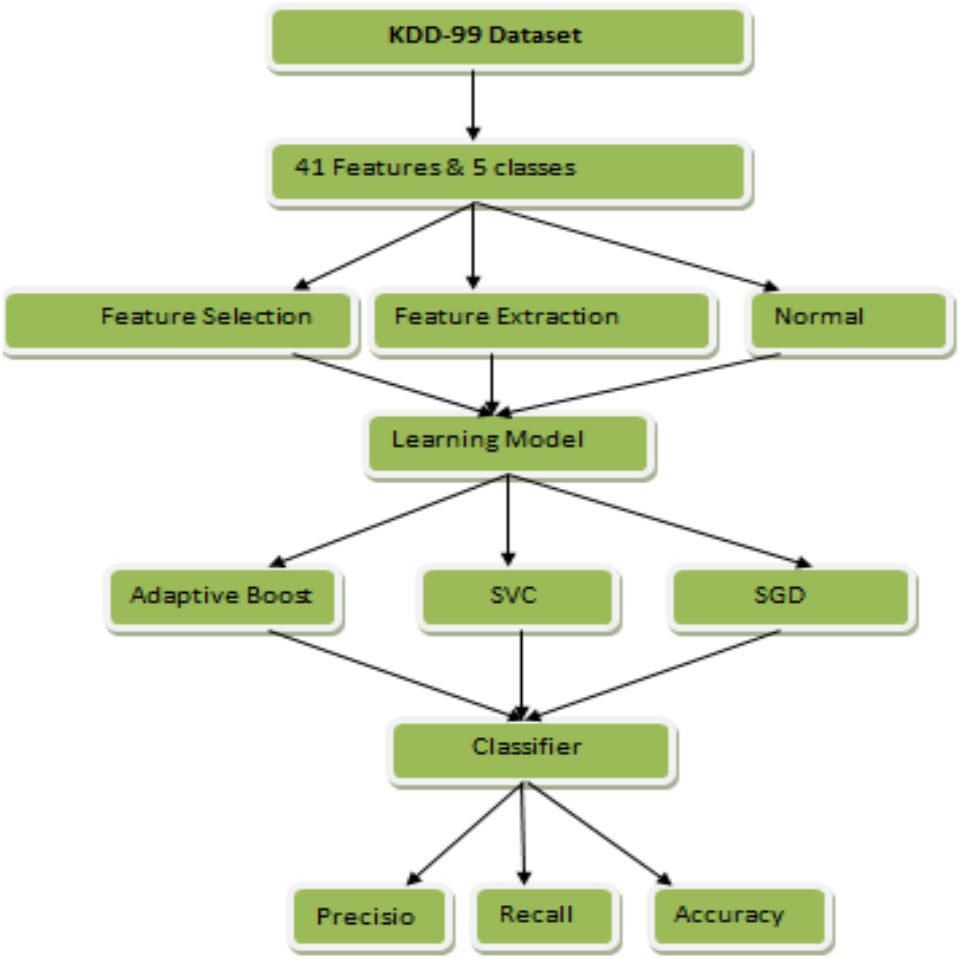
Recall: It is equivalent to the true positive rate (TPR). It measures performance of the algorithm.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Accuracy: It shows how accurate the system can detect attacks.

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

Where TP,FP,FN & TN are the numbers of True, Positives, False Positives, False negatives & True Negatives respectively.



V. RESULT

Main purpose is here to calculate the main metrics for the features extraction. The accuracy metrics are calculated with the help of a Machine Learning - Confusion Matrix.
Features instructions by KPCA:

Table 1. Forty one features comparison.

Classifier	Accuracy	Precision	Recall
Linear svm	90.91	84	86.182
SGD	96	96	96.182
ADA	84	84	86.182

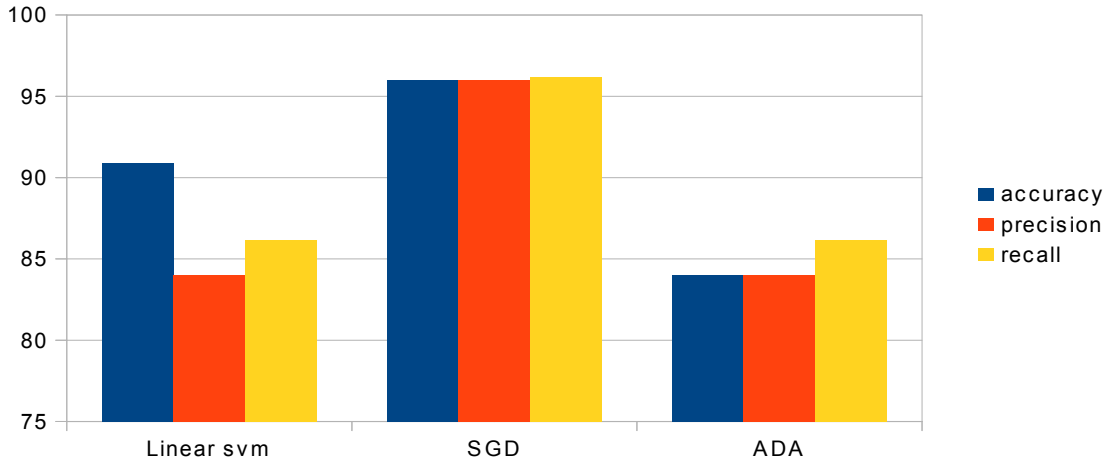


Fig1.Forty one features comparisons

Table2. Thirty five features comparisons

Classifier	Accuracy	Precision	Recall
Linear svm	90.91	73	62.182
SGD	86	84	76.182
ADA	71	96	96.182

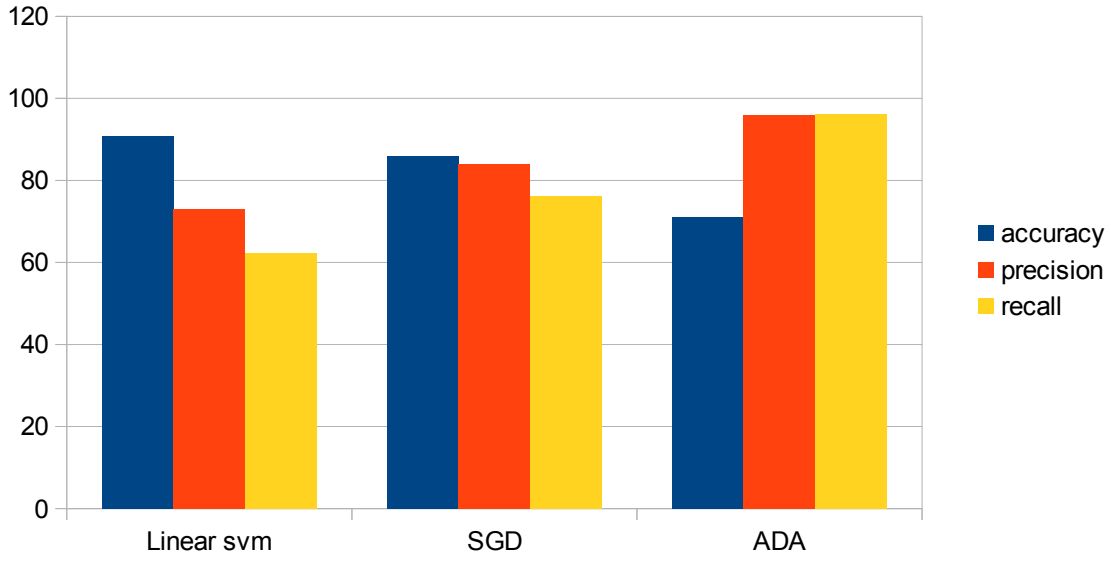


Fig 2.Thirty five features comparison

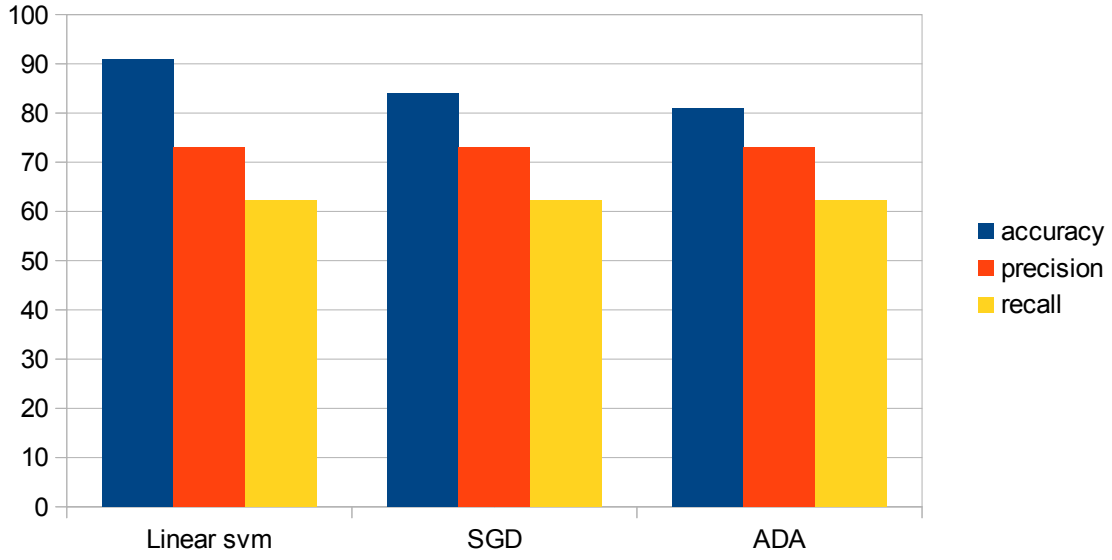


Table3. Twenty five features comparison

Classifier	Accuracy	Precision	Recall
Linear svm	90.91	73	62.182
SGD	84	73	62.182
ADA	81	73	62.182

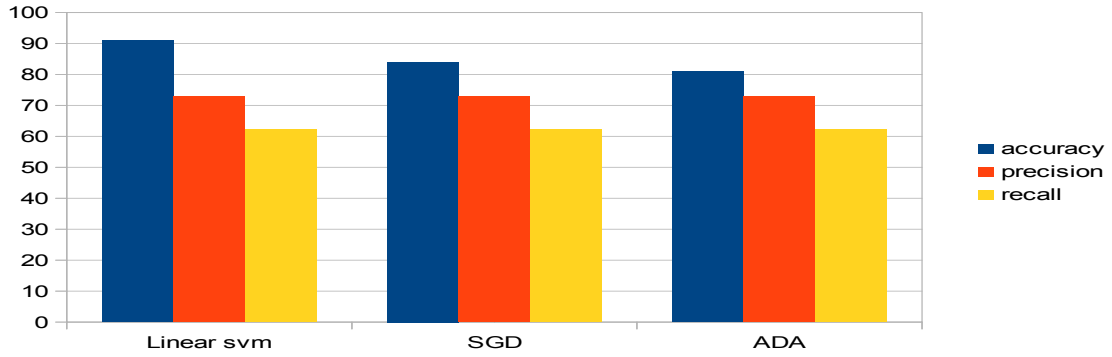


Fig3. Twenty five features comparison

Table4. Fifteen features comparison

Classifier	Accuracy	Precision	Recall
Linear svm	90.91	63	72.182
SGD	81.9998	64	62.182
ADA	75	82	61.23

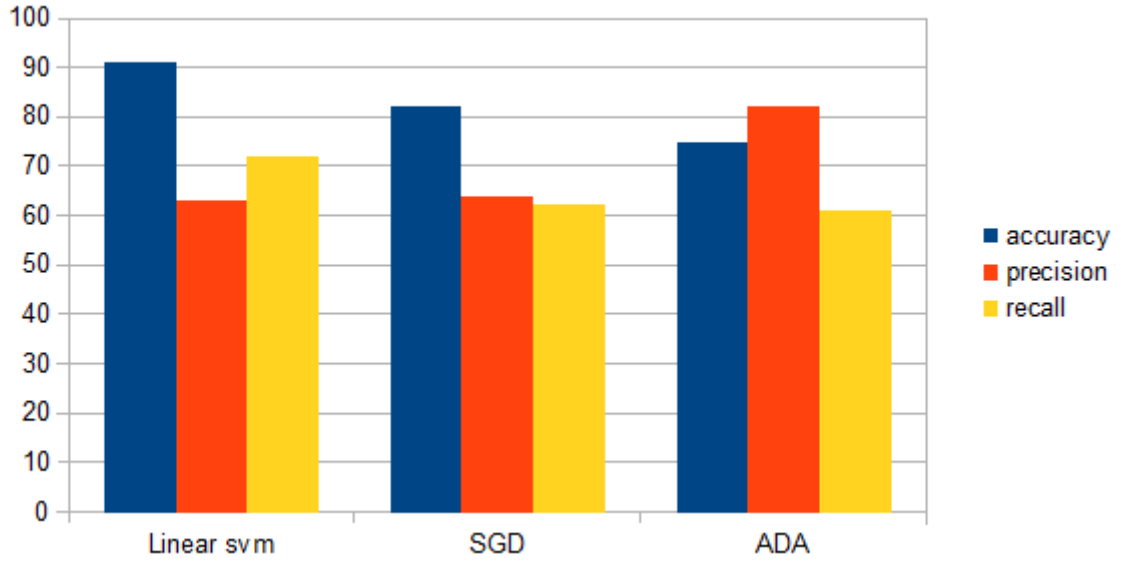


Fig4. Fifteen features comparison

Table5. Correlations instructions by KPCA:

SVM			
Features	**Accuracy**	**Precision**	**Recall**
41	0.9091	0.84	0.86182
15	0.9091	0.64	0.7418
25	0.9091	0.76	0.6618
35	0.9091	0.76	0.6618

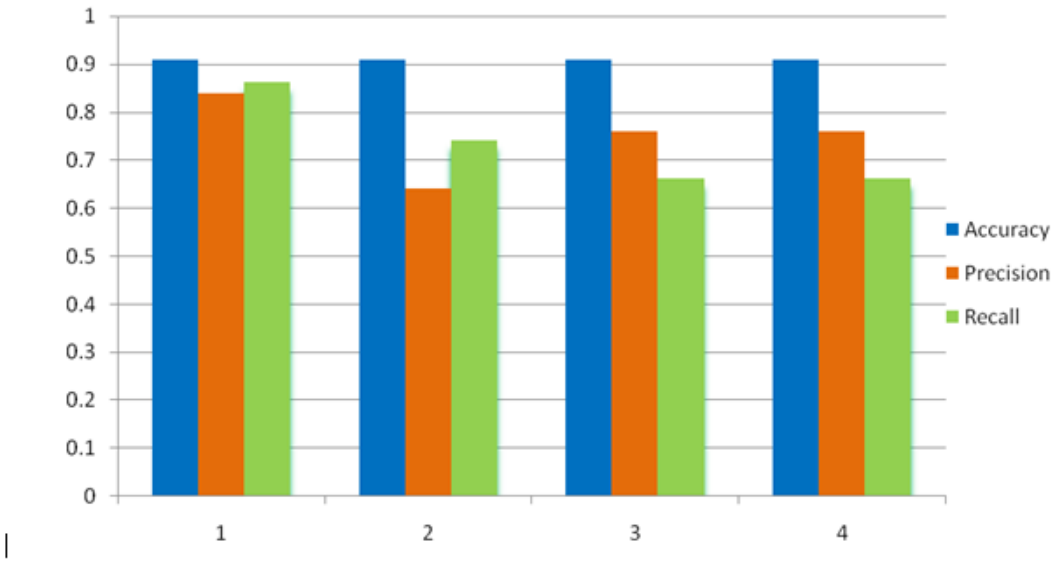


Figure 1 LINEAR SVM

Table6.

SGD			
Features	**Accuracy**	**Precision**	**Recall**
41	96	0.96	0.96182
15	80.9999	0.57	0.5818
25	85.9998	0.71	0.6818
35	84	0.82	0.7018

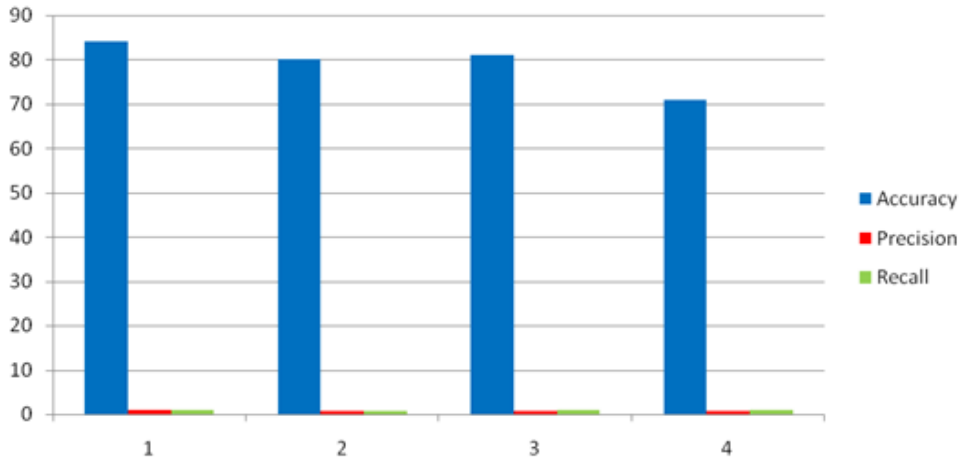
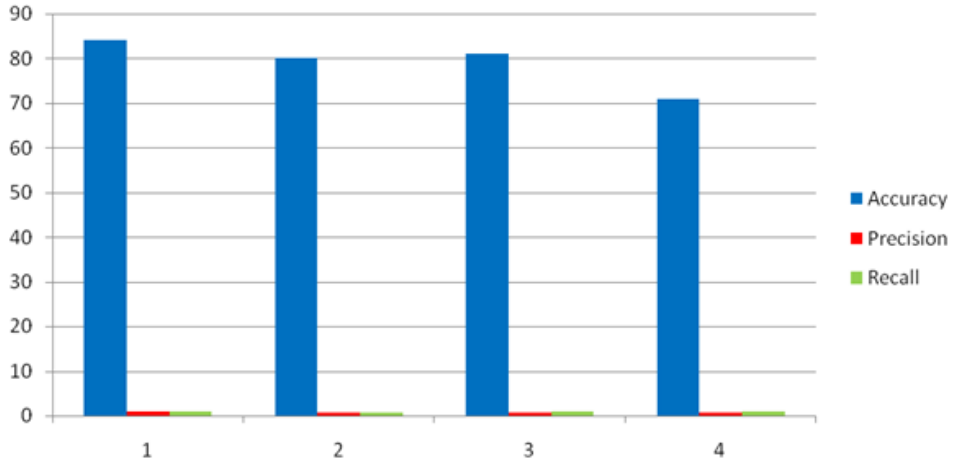


Table7.

Adaptive Boost			
Features	**Accuracy**	**Precision**	**Recall**
41	84	0.84	0.86182
15	80	0.74	0.82
25	81	0.76	0.91
35	71	0.6618	0.9018



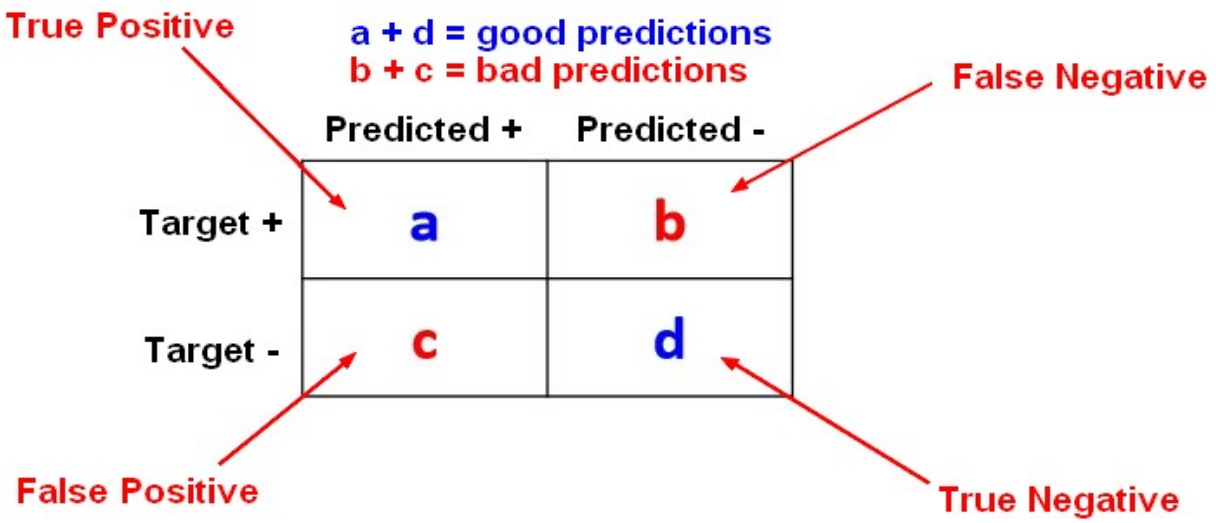


Figure 8: Confusion Matrix

1. Accuracy

Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. It may seem obvious that the ratio of correct predictions to cases should be a key metric. A predictive model may have high accuracy, but be useless.

Acc. = TP+TN / (TP+TN+FP+FN) where

TN is the number of true negative cases
FP is the number of false positive cases
FN is the number of false negative cases
TP is the number of true positive cases

2. Precision

Precision (P) is defined as the number of true positives (T_p) over the number of true positives plus the number of false positives (F_p).

Precision = TP / (TP+FP)

3. Recall

Recall (R) is defined as the number of true positives (T_p) over the number of true positives plus the number of false negatives (F_n).

Recall = TP / (TP+FN)

CONCLUSION

Experimental results shows that rare class predictive models on the KDDCup'99 dataset have much more efficiency in the detection of intrusive behaviour and DARPA 1998 data set experimental result shows that the new techniques have great promise in detecting novel intrusions.

REFERNCES

[1] Kumar J,D, "*Attack Development for Intrusion Detection Evaluation Attack Development for Intrusion Detection Evaluation*", Massachusetts Institute of Technology,2000.
[2] Chebrolu S., Abraham A., and Thomas P., "*Feature Deduction and Ensemble Design of Intrusion Detection Systems*," Computers and Security, vol. 24, no. 4, pp. 295-307, 2005.
[3] S.S.Garasia, D.P.Rana and R.G.Mehta, "HTTP BOTNET DETECTION USING FREQUENT PATTERNSET MINING", INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY, Volume-2, May-Jun 2012.
[4] Carl Livadas et al, "Using Machine Learning Techniques to Identify Botnet Traffic", IEEE, Local Computer Networks, Proceedings 2006 31st IEEE Conference on 14-16 Nov. 2006.
[5] Inam Mohammad et al, "A Review of types of Security Attacks and Malicious Software in Network Security", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.
[6] Lee, Wenke, and Salvatore J. Stolfo. "Data Mining Approaches for Intrusion Detection." *Usenix security*. 1998.

[7] Mukkamala, Srinivas, Guadalupe Janoski, and Andrew Sung. "Intrusion detection using neural networks and support vector machines." *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*. Vol. 2. IEEE, 2002.
[8] Barbara, Daniel, et al. "ADAM: Detecting intrusions by data mining." *In Proceedings of the IEEE Workshop on Information Assurance and Security*. 2001.
[9] Honig, Andrew, et al. "Adaptive model generation: architecture for deployment of data mining-based intrusion detection systems." *IN*. 2002.
[10] Dickerson, John E., and Julie A. Dickerson. "Fuzzy network profiling for intrusion detection." *Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American*. IEEE, 2000.

