International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1520-1527

Performance Analysis of Big Data Intrusion Detection System over Random Forest Algorithm

Alaa Abd Ali Hadi

Al-Furat Al-Awsat Technical University, Iraq.
alaaalihadi@gmail.com

Abstract

The Internet has grown rapidly in the last ten years. Consequently, the interconnection of computers and network devices has become so complex for monitoring that even the security experts do not fully understand its deepest inner workings. Personal computers have become very fast every year. It is not rare for a very ordinary person to connect to the Internet through 20 Mbs lines or faster. With this huge network data the network security has becomes very important for monitoring the data. The big challenge of developing big data in intrusion detection system is time of building system. In this paper, the preprocessing feature section method has been applied to generated subset of relevant features for building the model. The Random forest algorithm was applied to classify the network data. To increase the accuracy of Random forest algorithm the information gain method was used. The NSL-KDD standard data was employed to examine the performance of the proposed model. Various evaluation metrics have proposed to evaluate the proposed model. The empirical results of proposed model show that it is better in terms of performance measures. A comparative analysis of the results obtained of proposed model and different various existing algorithms is presented. The results show that the performance of the proposed model outperformed the performance of existing systems.

INTRODUCTION

Network security has become an extremely critical part of global infrastructure, Considering the fact that personal, ecommerce, banking and business data is being shared on computer networks, security has become one of the major aspects of Internet. One of the important fields on network security is Intrusion Detection System (IDS). To mitigate and prevent attackers form intrusion network and awareness of the attacks is challenging as it is faced by network security communities. Yearly numbers of new vulnerabilities are discovered. The security tool has faced more difficulty to automate detecting the new attacks. The intrusion detection system has become very significant and helpful for protecting the computer network from attacks. For example many organizations or companies round the world use firewalls as a defending measure to protect their secret network data from the public networks. The firewall can be used to secure the resources from inside the uses but the security of the resources cannot be obtain as truly hundred percent of security. Furthermore, the intrusion detection system is very important network security aspect which is used to protect network and detect adversarial on the activities in a network . The IDS tool is working based on the supposition that an attacks activity's signature is different from the signatures of normal activity. The intrusion detection system has two ways for detecting the attacks either signature-based detection or anomaly based detection. Signature-based analysis is used to detect against a database of recognized attack signatures to determine attack matches. Whereas the anomaly based detection use monitoring against a normal baseline, and can issue alerts based on abnormal behavior.

BIG DATA IN INTRUSION DETECTION SYSTEM

This section describes the comprehensive introductory background and the framework of big data challenges facing Intrusion Detection. With increasing development of technology, huge amount of digital data is being produced every day. Big data is a term when large amount of data face the challenge of being processed by traditional approaches. During the last few years various methods are developed to process. and manage the big data. Big data in network security is a key issue for many researchers for finding specific solution due to network security use to protect many more security breaches such as finances, industry, medicine, and other important aspects. IDS is one security solution for any act of control for any intrusion pattern or for the unknown packets an inner network. Due to the complexity of network data the big data techniques are very important to analyze the network patterns and finding out what has happened in the network? Such as the network data having many different structures and formatting and capturing network data from different source that make it difficult to analyze the same by using traditional approaches. Furthermore, the network data face big problems of high dimensionality. Machine Learning method is widely employed to help for building Intrusion Detection Systems from zero day attacks.

In this research work, it has mainly focused on the big data in Intrusion Detection over Machine Learning algorithms. The proposed method of Information with random forest approaches is used to detect intrusions from network big data. The motivation of this research work is the proposed new model which can help to

detect intrusion with more accuracy and faster. Furthermore, the feature selection methods are more important for dimensionality reduction data set. The complexity time and accuracy of a classifier is greatly reduced if the numbers of features of data set are reduced.

The paper is organized as follows section 1 is an introduction. The Big data in Intrusion Detection System is introduced in section 2. In section 3 presented related works. The Methodology of proposed model is presented in section 4. In section 5 experimental analysis is described. Performance comparison is presented in section 6. Finally, the paper is closed with conclusion in section 7.

RELATED WORK

Numbers of research outcomes in relation to intrusion detection system using data machine learning algorithm published in literature review are interested in solving the problem of improving the efficiency of intrusion detection system using machine learning approaches. Table 1 shows few of the reviews of existing research works on several machine learning algorithms that are employed for big data in Intrusion Detection System.

Table 1.1 Summary of related work

Authors	Research Statement	Methodology	Techniques	Main Work
Chitrakar et.al. [1]	Anomaly based Intrusion Detection	Experimental	K-Medoids clustering and Naïve Bayes classification	Reduce False Alarm Rate and increase detection rate
Dhakar et al. [2]	Hybrid Intrusion Detection Systems	Experimental	Tree Augmented Naïve Bayes and Reduced Error Pruning	Able to detect unknown intrusions and reduced the false alarm rate
Huai-bin et al. [3]	Clustering algorithm for intrusion detection system	Experimental	SOM neural network and K-means Algorithm	High detection rate and controls the false positive rate in the low range
Marcelo et al.[13]	Map reduce for intrusion detection	Experimental	Map reduce method	Time reduction
Lidong Wang et al.[14]	Big data review of intrusion detection system	Review	Reviewed different techniques of intrusion detection system	Shown challenges big data of intrusion detection
Jingwei Huang et al.[15]	Classified bigdata intrusion detection system	Experimental	LDA	Detection rate
Rachana Sharma et al.[16]	Classified bigdata intrusion detection system	Experimental	K-Nearest Neighbor	Performance of Mapreduce with KNN classifires Detection rate and FPR
Miss Gurpreet et al.[17]	Big Data for Detecting Unknown Attacks	Experimental	Pattern matching methods	Detection unknown attack

METHODOLOGY

An analysis of big data in intrusion detection system is the main objective of the present research work. Figure 1 displays the proposed model of big data intrusion detection system. An experimental study is carried out with use standard intrusion detection system data set includes normal packets and abnormal packets. The data is preprocessed by applied information entropy method. The information entropy method is applied to build the classifiers with more accurate and fast. The Random forest algorithm is used to classify the data as normal and malicious packets. The performance measures are used to evaluate the classifier results. Finally a comparative results analysis between the proposed with different existing classifiers is presented. The detailed description of each step used in the proposed model is presented in the subsequent subsection.

NSL-KDD data set

The NSL-KDD data is used to test the IDS proposed model. NSL-KDD standard intrusion detection system data set is an updated version of KDD cup'99 data set. Due to the problem of KDD cup 1999, the NSL-KDD data set is developed to solve the problem discussed by McHugh [8]. To run the experiments on the complete set without the need to randomly select a small portion. The NSL-KDD data set contains 4,898,431 entries. The NSL-KDD data set are collected as raw network packets and are independent of an operating system or an application. Consequently, each record in this data set has been provided a label recognizing to which class label the record belongs to. All labels in this data set are supposed to be correct. The NSL-KDD contains 37 attack types. The simulated attacks fell in precisely one of the four categories: Denial of Service, Probe, User to Root and Remote to Local. Table shows all types of attacks in NSL-KDD data set.

Table 2. All types of attacks in NSL-KDD

Attacks in Dataset	Type of attacks
Dos	Back, Land, Neptune, Pod, Smurf, Teardrop, Mailbomb, Processtable,Udpstor m,Apache2,Worm
Probe	Satan, IPsweep, Nmap, Portsweep, Mscan,Sa int
R2L	Guess_password,Ftp_write,Imap, Phf,Multihop,Warezmaster Xlock,Xsnoop,Snmpgue ss,Snmpgetattack, Httptunnel, Sendmail, Named
U2R	Buffer_ overflow, Loadmodule Rootkit, Perl,Sqlattack, Xterm, Ps

Preprocessing

Preprocessing is an important stage employed to control real world datasets into an comprehensible format. Certainly, the real world datasets have been incomplete, noisy in specific behavior. Preprocessing phase is very significant to analyze patterns from big data. Hence, the preprocessing methods are necessary in big data intrusion detection system to enhance the machine learning algorithm for classification of the patterns. Thus, supplementary to improve the precision and efficiency of resulting machine learning job. In the present research study the information gain method is used to obtain significant features from data set. The detailed description of information gain method is presented in the subsequent subsections.

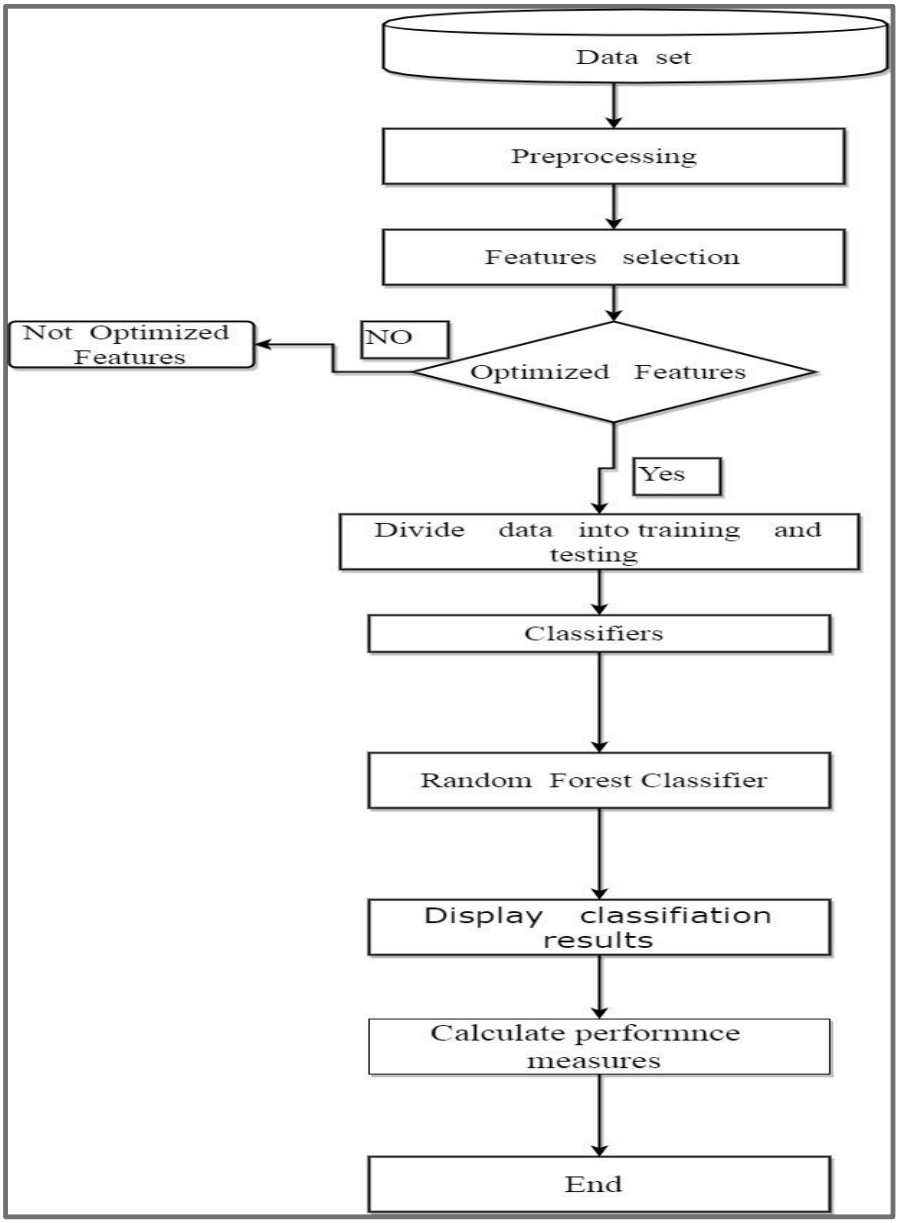

Figure 1. Proposed model

Information Gain (IG) method

Information gain is a preprocessing step. Information Gain is used to evaluate the worth of the features by measuring entropy with respect of class. The entropy theory comes from the theory Which higher the entropy of features indicates to the more the information content. The information gain uses the concept of entropy which selects the subset features which have highest information rank from data set. Due to this feature it has more power to classify the data. Usually, the information gain is defined as a joint set of features as the reduction in the entropy that is archived by learning a joint feature set *F*.

$$Entropy(s) = Info(G) = -\sum_{i=1}^{m} p_i \quad log_2(p_{i)} \qquad (1)$$

Where G denote to the probability of occurrences of the class label over the total of class label data set. Where p_i is random probability that belongs to the class lab C_I. In this presented work the information Gain method has used to select the most significant features from data set. When applying the information gain method, 13 features have selected that are most significant out of 41 features which are listed in table 2. Figure 2 displays the procedure of generated the subset features

Table. 3. 13 significant attributes obtained by an information gain method

Feature set	Rank
service	1.576691
flag	1.283756
src_bytes	0.986585
dst_bytes	0.983083
count	0.965245
serror_rate	0.953428
same_srv_rate	0.8945
diff_srv_rate	0.892471
dst_host_srv_count	0.878602
dst_host_diff_srv_rate	0.861297
dst_host_serror_rate	0.676338
dst_host_srv_serror_rate	0.61901

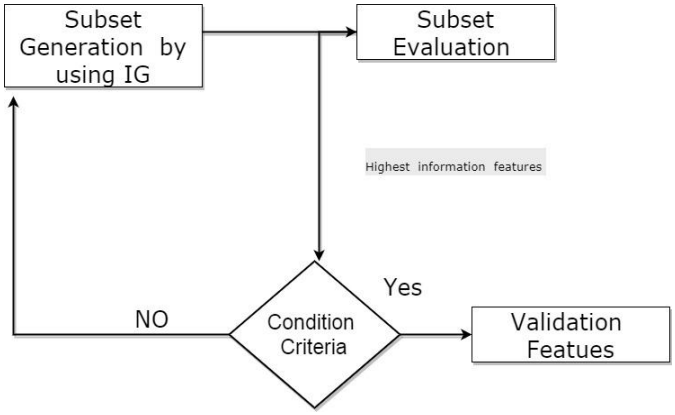

Figure .2 Procedure of selection the features

Random Forest algorithm

Random Forest algorithm is one of a supervised classification algorithms. Form the name of the algorithm is working as create the forest in randomly in proper way. The Random Forest is decision tree algorithm which uses to building several trees then joining their output to enhance generalization ability of the proposed model. Random Forest algorithm is Ensemble of combining several individual trees to produce a strong learner tree. Random forest algorithm generates several classification trees to obtain one stranger leaner tree. Random Forest is working to construct each tree by using a different bootstrap sample from the original data set using a tree classification algorithm. When the forest is constructed, a new object that requires to be classified is placed down at every tree in the forest for the purpose of classification purpose. Finally each tree in forest has a decision to select the class of object. The Random Forest algorithm has been used to develop IDS system.

Performance measures

The performance measures have been carried out to test the results of proposed model. The Accuracy, False Positive, Precision, True Positive and Time are used. The equations performance measures are as follow:

$$\text{False Positive Rate (FPR)} = \frac{FP}{TN+Fp} \% \ 100 \qquad (2)$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN} \%100 \qquad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \%100 \qquad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \%100 \qquad (5)$$

True negative (TN): Correctly classified of valid records as normal record.

True positive (TP): Correctly classified of attack records as attacks.

False positive (FP): The percentage of incorrect records normal data as attacks.

False negative (FN): The percentage of incorrect records attacks as normal record.

EXPERIMENTAL SETUP

The proposed model big data intrusion detection system method is implemented by using MATLAB R2013a-64 windows 7 Ultimate with the core i5 processor and 8 GB RAM , and the Weka tool was used to compare with existing different classification algorithms. Various performance measuring have been used to test the proposed model. In this experiment hybrid model of random forest and information gain algorithms are applied. In this experiment, 31 major attacks are selected. The data has only 185559 attacks and normal instances are used. These attacks correspond to 185559 in total data set. The original data set contains 25 MB data. The hybrid model of Random forest and information gain algorithms are applied on Matlab. Table 4 shows the performance of the proposed model . It investigated that the correct classification of instance is 184331 out 185559 instances. Furthermore, 1228 instances is misclassification out of 185559 instances.

Table 4: Performance analysis of proposed model

Performance	Proposed model
Time	16.87 seconds
Correctly Classified Instances	184331
Incorrectly Classified Instances	1228
Total Number of Instances	185559

In order to enhance the proposed model, it is decided to work with preprocessing. Feature selection method. The information gain method is proposed for improving the Random forest classifier. The feature selection method help to increase the accuracy of classifier and reduce the time of building the model. After obtaining the goodness features, it was challenged by selecting the number of subset of relevant features from original data set. Finally, the features have highest rank of information have been selected for enhancing the classification accuracy. Table 5 shows performance of different existing classifier against the proposed model with using feature selection method. It is observed that the proposed model has outperformed better than the all existing algorithms.

Table 5: Results of proposed model with different existing algorithms

Classifiers	FP	TP	Accuracy	Precision
Naïve Bayes	0.003	0.949	94.9261	0.949
REP Tree	0.003	0.988	98.767	0.721
SVM	0.004	0.957	95.43	0.987
KNN	0.007	0.933	93.12	0.975
Proposed model	**0.001**	**0.993**	**99.33**	**0.993**

PERFORMANCE AND COMPARISON OF PROPOSED MODEL

The Comparison standard to evaluate and examine the proposed model of intrusion detection system with respect to classification accuracy of intrusion detection system . The FP, TP, accuracy and precision performance measures are used to investigate the proposed with all various existing algorithms. Table 4 summarizes the obtained results of the proposed and existing algorithm with using features selection method. It is observed that the results of the proposed model is better than the all existing algorithms. Figure 3 illustrates accuracy results performance of the proposed model in comparison with different existing algorithms. It shows that the proposed model is more accurate and takes less time to build model. Figure 4 displays the false positive performance of the proposed model and existing classification algorithms, it shows that the FP of the proposed model is very less. The TP and precision measures of the proposed model is the highest with in comparison with another existing classifiers that is displayed in figure 5. Finally, it is concluded that the proposed model can detect different types of attacks with best accuracy compared to other existing systems.

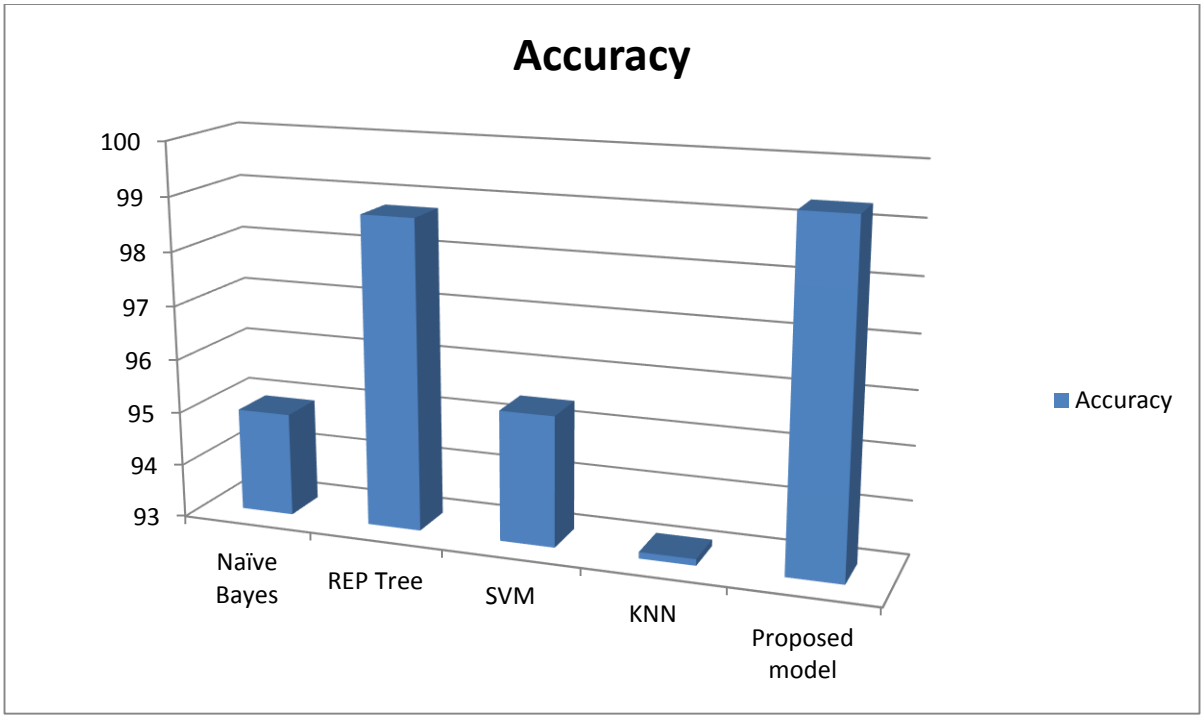

Figure 3 performance accuracy of proposed model and existing models

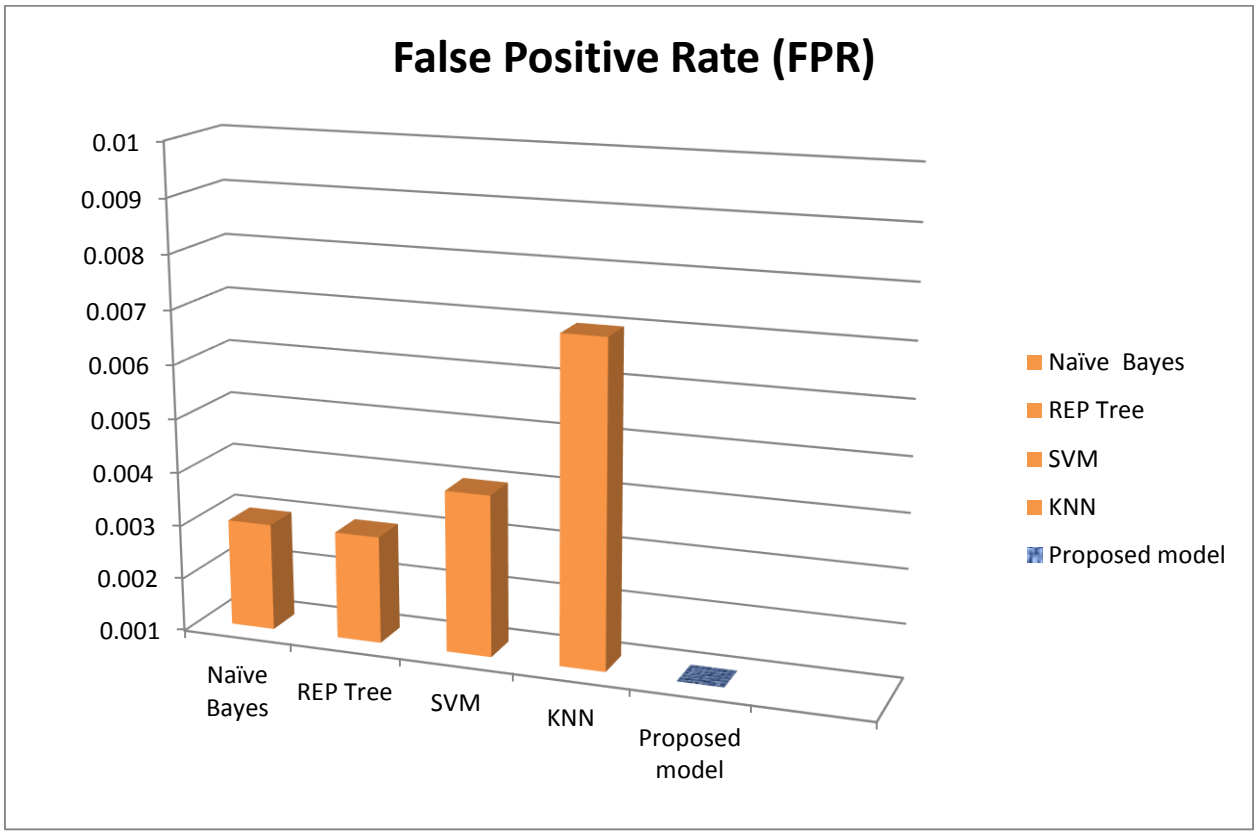

Figure 4 performance FP of proposed model and existing models

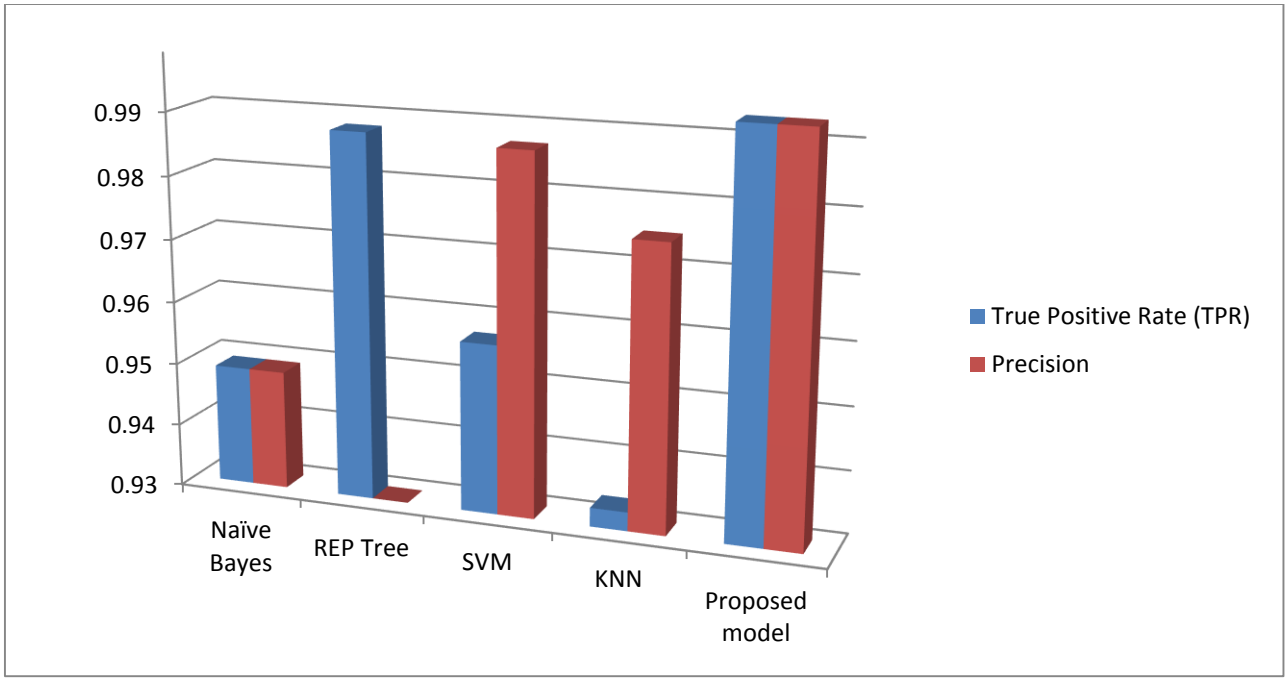

Figure 5 performance TP and precision of proposed model and existing models

CONCLUSION

The goal of the research is to improve the existing algorithms for building Intrusion detection system. It has succeeded in achieving the main target by using proposed model. The proposed of Random and information gain methods have enhanced the accuracy of big data in intrusion detection system. In this research all types of attacks have used. Due to the big data the time is taken to build system of IDS. However, the features selection method is applied to solve the problem. After Information gain is used the 13 most significant subset features from the original 41 features from data is generated. Furthermore, the proposed model is applied using these most significant features. It is observed that the accuracy is increased and the time of building the model is reduced. The various performance measures were employed to test the proposed model. The accuracy of proposed model is 99.33%. A comparative results analysis between the proposed model and different existing model is presented. It is observed that the proposed model has outperformed the existing classifiers. In future, the researcher will try to use soft computing by using different data sets.

REFERENCE

[1] R. Chitrakar, and C. Huang, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-Medoids Clustering and Naïve Bayes Classification," In Wireless Communications, Networking and Mobile Computing (WiCOM), 8th International Conference on, pp. 1-5, IEEE, 2012.

[2] M. Dhakar, and A. Tiwari, " A novel data mining based hybrid intrusion detection framework," Journal of Information and Computing Science, vol 9, no. 1, pp. 037-048, 2014.

[3] W. Huai-bin, Y. Hong-liang, X. U. Zhi-Jian, and Y. Zheng, "A clustering algorithm use SOM and K-means in intrusiondetection," In E-Business and E-Government (ICEE), 2010 International Conference on, pp. 1281-1284, IEEE, 2010.

[4] S. Warnars, "Mining Patterns with Attribute Oriented Induction," In Proceeding of The International Conference on Database, Data Warehouse, Data Mining and Big Data (DDDMBD2015), pp. 11-21, 2015.

[5] V. Kachitvichyanukul, "Comparison of three evolutionary algorithms: GA, PSO, and DE," Industrial Engineering andManagement Systems,11(3), pp. 215-223, 2012.

[6] R. Chitrakar, and C. Huang, "Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-MedoidsClustering and Naïve Bayes Classification," In Wireless Communications, Networking and Mobile Computing (WiCOM),8th International Conference on, pp. 1-5, IEEE, 2012.

[7] Shi, X., Manduchi, R., 2003. A study on Bayes feature fusion for image classification. In: Conference on Computer Vision and Pattern Recognition Workshop, CVPRW, Madison, Wisconsin, USA, pp. 95–95.

[8] http://www.kdd.ics.uci.edu/databases/kddcup99/task.html 7

[9] Nassar M, al Bouna B, Malluhi Q (2013) Secure

outsourcing of network flow data analysis. In: Big Data (BigData Congress), 2013 IEEE International Congress On. IEEE, Santa Clara, CA, USA. pp 431–432

[10] Kezunovic M, Xie L, Grijalva S (2013) The role of big data in improving power system operation and protection. In: Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium. IEEE, Rethymno, Greece. pp 1–9

[11] Denning DE (1987) An intrusion-detection model. Softw Eng IEEE Trans SE-13(2):222–232. doi:10.1109/TSE.1987.232894

[12] Suthaharan S, Panchagnula T (2012) Relevance feature selection with data cleaning for intrusion detection system. In: Southeastcon, 2012 Proceedings of IEEE. IEEE, Orlando, FL, USA. pp 1–6

[13] Marcelo D. Holtz, Bernardo M. David and Rafael Timeote "Building Scalable Distribute Intrusion Detection System Based on the Map Reduce Framework. 2011, Intrenation journal of Revista Telecommucation pp 23-31

[14] Lidong Wang*, Randy Jones "Big Data Analytics for Network Intrusion Detection: A Survey. International Journal of Networks and Communications 2017, 7(1): 24-31 DOI: 10.5923/j.ijnc.20170701.03

[15] Jingwei Huang, Zbigniew Kalbarczyk, and David M. Nicol. "Knowledge Discovery from Big Data for Intrusion Detection Using LDA. 2014 IEEE International Congress on Big Dat pp760-762

[16] Rachana Sharma & Priyanka Sharma, Preeti Mishra & Emmanuel S. Pilli "Towards MapReduce Based Classification approaches for Intrusion Detection". Intrnation conference 2016 IEEE PP-361-366

[17] Miss Gurpreet Kaur Jangla1, Mrs Deepa.A.Amne2.." Development of an Intrusion Detection System based on Big Data for Detecting Unknown Attacks. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015