Performance Assessment of Ml Techniques for Detecting Intrusions in the Network

Katikela Haritha, CH. Mallikarjuna Rao

Abstract : It has become crucial for the organizations, military and personal computer users to make the network security. Day by day, security has become a major issue with the increase of internet usage. The improvement in the security technology can be much understood from the security history. Network security is an immense field and it is in development stage. An immense amount of data is being generated every second due to technological advancement and reforms. Social networking and cloud computing are generating a huge amount of data every second. Every minute data is being captured in the computing world from the click of the mouse to video people tend to watch generating an immediate recommendation. Everything a user is doing on the internet is being captured in different ways for multiple intents. Now it all ends up monitoring the system and network and, securing lines and servers. This mechanism is called Intrusion Detection System(IDS). Hacker uses multiple numbers of ways to attack the system which can be detected through a number of algorithm and techniques. A comprehensive survey of some major techniques of machine learning implemented for detecting intrusions. Classification techniques are SVM, Random Forest algorithm, Extreme learning machine, and Decision Tree. NSL-KDD is the dataset used to get the higher rate of detection. The Result Analysis shows that, in terms of accuracy, this paper accomplishes better results when compared to any other related methods.

Keywords: Detection rate, Decision Tree, ELM, Machine Learning (ML), NSL_KDD dataset, Random Forest, Support Vector Machine.

I. INTRODUCTION

Internet usage not only has advantages but also created different ways to compromise system stability and system security connecting to it. Even though static defense mechanisms like software updates and firewalls gives us certain level of security, IDS should also be used as it is a dynamic defense mechanism.

Intrusion detection is a process of determining the signs of intrusions by observing the events appearing in a computer system/computer network. IDS are segregated as host based or network based. Network-based analyzes for the signs of intrusions from gathered raw network packets while Host based functions on the information gathered from personal computer system for analyzing the signs of intrusions. In order to find the attack patterns, two different detection techniques were employed in Intrusion Detection System.

Revised Manuscript Received on October 05, 2019.

Katikela Haritha, PG Scholar, M.Tech, Department of Computer Science and Engineering, GRIET , Affiliated to JNTUH, Hyderabad, India. Email: haritha.katikela@gmail.com

Dr. CH. Mallikarjuna Rao, Professor of CSE, GRIET, Affiliated to JNTUH, Hyderabad, India. Email:chmksharma@yahoo.com

They are Anomaly and Misuse detection systems. Finding the attacks by identifying the alteration in patterns of utilization or system behavior is termed as Anomaly detection system [1]. Whereas, in the monitored resources searching for the known attack signatures is termed as Misuse detection system. Besides the security infrastructure of most of the organizations, IDS has become crucial as the network attacks are increasing over the past few years [2]. As there is no possibility to arrange a system theoretically without vulnerabilities, it is an extreme challenge to deploy a very effective Intrusion Detection System and it appears to be an important in the research field [3]. For detection of intrusive activities from the dynamic and complex datasets, various machine learning (ML) algorithms have been used such as Fuzzy Logic [3][4], Genetic Algorithm [4][6], Clustering Algorithm[5], Neural Network [6], etc.

In order to provide a better solution for the IDS problem, recent time RF and SVM have been used. Performance analysis of 2 commonly used Machine Learning techniques is concentrated in this paper. They are SVM and Random Forests (RF). To train and test the model, KDD99 dataset has been used.

The classification of the model is then analyzed in terms of efficiency and effectiveness. Accuracy remains an issue, eventhough the availability of various ID techniques. The accuracy depends on rate of false alarm and rate of detection. This problem can be solved by improving the rate of detection and decreasing rate of false alarms. For the research work, this approach would be a motivation.

So, Random Forest(RF), support vector machine (SVM), Extreme Learning Machine (ELM) and Decision Tree(DT) algorithms are used in this work. To solve this classification problem, these techniques have been determined to be effective in their capability.

ID mechanisms are verified on an approved KDD dataset. This work used an improved form of KDD dataset i.e., NSL_KDD dataset, which is treated as standard dataset in the estimation of ID methods.

II. RELATED WORK

As compromised information cause damage to individuals and organizations, it is important to secure the network and computer information. Intrusion detection systems are crucial to prevent such circumstances. In order to enhance the IDS performance, in the recent work various ML techniques and data mining techniques have been suggested.

Wang *et al.* [7] recommended SVM based Intrusion detection framework and evaluated on NSL_KDD dataset. They declared that other approaches has less rate of effectiveness when compared to their suggested approach. But the size of testing and training sample of dataset is not mentioned by them. So, its not good opinion to use SVM to evaluate the large network data for intrusion detection, as performance of SVM reduces when dealing with huge data.

Detection, Prevention and Resistance of Unauthorized access can be provided by IDS systems. Hence, Reaz and Aburomman [8] suggested an ensemble classifier approach i.e., a combination of SVM and PSO. This obtained a higher accuracy of 92.90 percent than the other approaches. KDD99 is the dataset used by them and it has the drawbacks as mention earlier. Moreover it's not a best option to use SVM for processing large data as performance decreases when size of dataset increases.

Kuang *et al.* [9] suggested a model for intrusion detection and obtained a detection rate of 96 percent. The suggested approach is a hybrid approach of KPCA and SVM including GA for detecting intrusions. The dataset used by them for evaluation is the KDD CUP99. But this dataset has some of the constraints like redundancy. Due to redundancy, classifier may get biased to more repeatedly appearing records. For the Feature reduction process, they used KPCA. From the principal space ,because of choosing the principal component's top percentages, KPCA has some limitations like chance of lacking some useful features. Further, it's not a better option to use SVM for monitoring the large network data.

The most demanding problem in our daily life is to provide the security for the Network systems. As a prime defense technique, IDS is important. Hence, Zhu and Teng [10] implemented a SVM and Decision Trees based model and used the KDD CUP99 dataset. They tested on this dataset and obtained an accuracy of 89.02 percent. For the huge datasets, due to low performance and high computation cost, SVM is not preferred.

Raman *et al.* [11] suggested hypergraph- genetic algorithm(HG-GA) based ID system for feature selection and parameter setting in Support Vector Machine. They used NSL-KDD dataset and declared that they obtained a higher detection rate than the existing approaches.

Elbasiony *et al.* [12] suggested RF and Weighted KMeans-based-ID model & it is tested on KDD 99 dataset. For predicting the real time traffic, Random Forest is not an appropriate choice due to its slowness, as it forms many trees. Furthermore, this dataset has some limitations.

Jabbar and Farnaaz [13] proposed a Random Forest based IDS model. They used NSL-KDD dataset and tested the effectiveness of their model with this dataset and obtained a detection rate when compared to J48.One of the constraint of RF algorithm is that it makes the real time predictions slower as it forms many trees.

III. PROBLEM DEFINITION

NSL-KDD dataset is statistically analyzed and observed having some of the problems which effects the evaluated system performance and obtains a poor evaluation of anomaly_detection approaches. This dataset possess many redundant records and it's one of the major drawback.

Examining the test and train datasets of NSL_KDD, Mohbod Tavallaee identified that the records were replicated about 75 percent in test sets and 78 percent in train sets. Learning algorithms will get biased towards the higher frequent records due to the train set containing more redundant records.

Hence it causes more harm to networks like U2R attacks as it restricts from learning unfrequent records.

Due to the test set containing repeated records, causes the evaluation results to be biased by the methods which have better detection rates on repeated records. For the classifiers to not to get biased, by removing the duplicate records from NSL_KDD test and train dataset, we have derived a new dataset (10 percent NSL-KDD and corrected NSL-KDD). Without the need of selecting the small random portion of data, it makes us affordable to run the experiment with whole data as there are reasonable number of train and test records.

IV. IMPLEMENTATION METHODOLOGY

The proposed model has four steps. They include Collection of the Dataset, Preprocessing step, Classification step and the final step is Result Evaluation. Performance of the proposed system is greatly influenced by all these four steps. The main concentration is to examine the performance of various classifiers for detecting intrusions. They are Random Forest, SVM , ELM and Decision Tree classifiers. Proposed Design of the Intrusion Detection System is shown in fig:1.

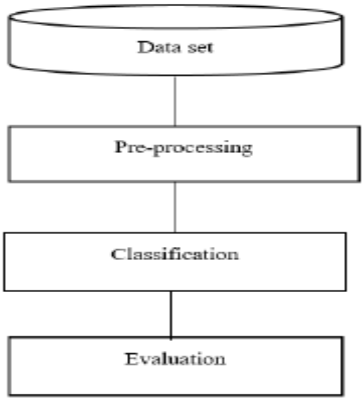

Fig. 1. Design of Intrusion Detection System.

A. Dataset

Technology is rapidly changing day by day, and hence a lot of inventions and technology advancements are being done in order to protect Computer Systems from any network intrusions attacks. Research on the network intrusion attacks, usually used the available dataset i.e., KDD CUP 99 and they applied various machine learning algorithms on this dataset. But this dataset was seen of having many problems.

Firstly, possess the repeated records in the KDD dataset is one of its major disadvantages. 78 percent of records are duplicated in the training dataset whereas, 75 percent of records are duplicated in the test dataset, and hence shifting our results on the learning algorithms to come out as biased.

Though, this dataset is publically available for detecting the network intrusions, it has some constraints. NSL_KDD dataset is the newly available version of this dataset. Now researchers are using the dataset to apply at the machine learning algorithms to it.

The new NSL KDD test and train dataset combines only very selected data records from the original KDD dataset and the redundancy of records do not exist. Hence, it being declared as a standard dataset and the research evaluation results be consistent over all researches. The training dataset is made up of basically 4 kinds of Attack Classes. They are , first is the Denial of Service (DOS)attack, second is called Probe Attack, Third is termed as User to Root (U2R) attack and the last class is Root to Local(R2L) attack. This in turn is made up of more than 21 different attacks.

B. Preprocessing

Due to the presence of some of the symbolic features in the dataset, classifier will not able to process it. As these symbolic features or non-numeric features do not have crucial participation in detecting intrusions, it is important to remove them. This process is termed as Preprocessing. Still, it makes an overhead involving higher training time. Classification architecture turn into complicated and it destructs the computing and memory resources. Hence, for improving the performance of IDS, symbolic features are removed from the raw dataset.

C. Feature Reduction

The main target of this work is to classify in a better manner. This is examined by deciding and examine to get higher accuracy in attacks classification and the time of training in NSL_KDD dataset. Correct way to classify the four kinds of attacks(DOS, Probe, U2R, and R2L) can also be tried to learn. In order to reduce the features, many of researchers have used different approaches. Knowing the actual amount of information present in the various features of the dataset could be the basic and clear approach. Computation of maximum entropy helped in reducing the number of features in the work[14].

Using 'k' different values, entropy set can be computed:

entropy(set)=I(set)=$\Sigma P(valuei)ki=$

1.log2$P(valuei)$

The formula represents getting probability of i th value as P(value i).

Primarily, we start examining the whole features, after then we compare the information gain by slowly decreasing the features count. It's been observed that alteration in information gain with whole features and with 18 to 20 features is relatively similar others are dissimilar.

D. Classification

The main objective of IDS is to classify the activity into intrusive & normal type and is also called as Intrusive Analysis Engine. Hence, for detecting the intrusive activities, many classifiers have been used like SVM, Self organizing map, multilayer perceptron and Naïve Bayes. Despite, based on their proven ability in classification four classifiers are used in this work .They are RF, ELM, SVM and DT. All the four classification algorithms details are given.

a) Support Vector Machine Model

To identify the better model for Support Vector Machine with different kernel, Grid search method has been used in this work. By assessing the various combinations of probable values, this approach opts a better outcome. Parameter range is considered in the grid search.

When compared with the other classifiers, the results noticed with the SVM model are unsatisfiable. Requiring minimum parameter adjustment is the benefit of SVM. It also has flaws like, in case classification for each of the instance in the training set, it requires gaussian function. With that, on the huge dataset with 1000s of instances the performance decreases and the training time expands. Soft margin is used if in case maximum margin classifier fails to identify the splitting hyper plane[15].

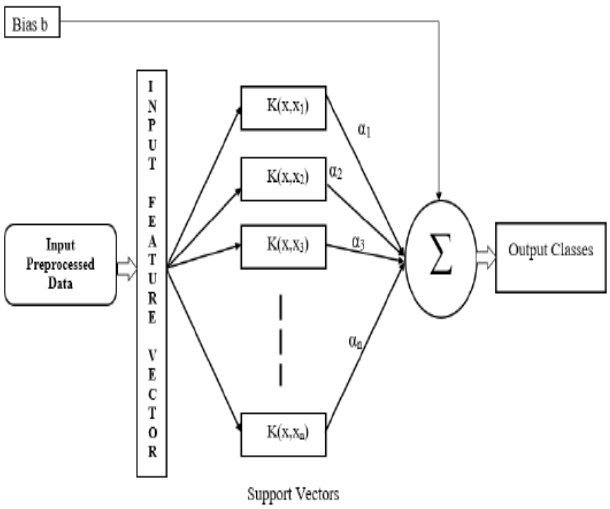

Fig. 2. Support Vector Machine Design for the detection of intrusions.

b) Random Forest Parameter Tuning

We optimize the random features count (mtry), to increase the rate of detection. Over the train set, with different mtry (5,6,7,10,15,20) we form the forest [16].

Then after , OOB error rate and the time taken to form pattern equivalent to different mtry is ploted. When mtry is 7, the OOB error rate attains minimum. In addition, if mtry is increased, the time taken to form random forest also increases. Hence, 7 is taken as an optimal value as it obtains a minimum OOB error rate and among all the values it takes minimum time.

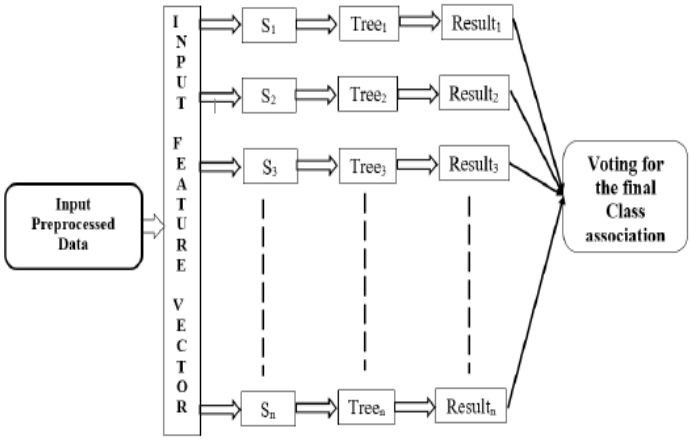

Fig. 3. Random Forest Design for the detection of intrusions.

In Random Forest, minimum size of the node and trees count are the 2 other parameters.

In RF as trees count (ntree) increases, trainingset is used for training and testingset is used for testing & the OOB error rate is examined with the testingset error rate to get the trees count. If there are enough trees count(around 100) , then the plot shows OOBerror rate tracks the testingset error rate closely. After the training error rate reaches 0, the testing set error rate and OOB error rate do not increase; this is an important characteristic of RF. Rather they converge close to their minimum i.e., their "asymptotic" values.

c) Extreme Learning Machine

Single or Multiple hidden layer feed_forward neural_network which is known as Extreme Learning Machine. Clustering, feature engineering, classification and the regression problems can be solved by ELM. It contains three layers. They are the input, 1 or more hidden and output layers.

As Traditional neural networks requires multiple rounds to converge, the process of adjusting the weights of hidden & input layer is expensive and time taken. Inorder to solve this issue, by arbitrarily selecting the hidden layer biases and input weights Huang et al. [17] suggested SLFN to reduce the time of training. Huang *et al.* [17] provides the complete details of ELM.

Many of the researchers determined that, when compared to the other feedforward networks, ELM attains higher generalization capability and learns faster.

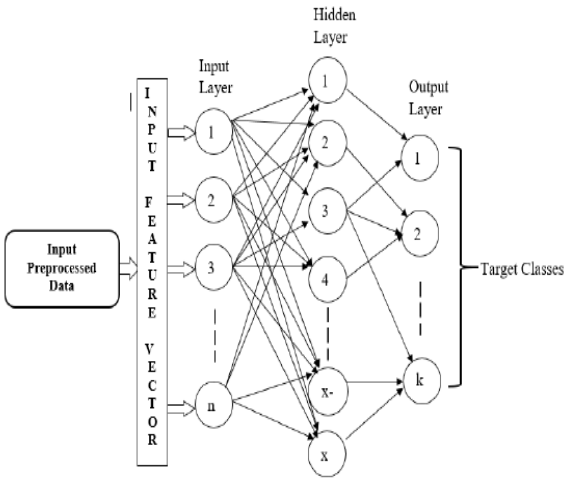

Fig. 4. ELM Design for the detection of intrusions.

d) Decision tree

It is treated as a classification problem for detection of intrusions. Based on the existing data, each and every connection is analyzed as normal type or the attack type. This problem of detection can be solved by DTs where it learns from the existing dataset and classifies the new data into one of the classes accordingly as specified in the dataset. As DTs learn a model based on the trained data and predict the future data based on the learned data into one of the type normal or attack class, it can be used for the misuse ID.

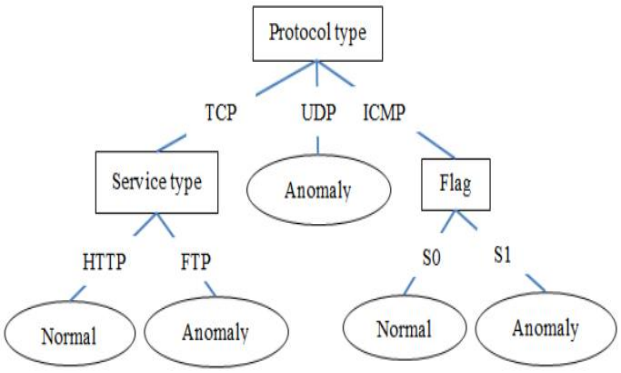

Fig. 5. Decision tree Design for the detection of intrusions.

When dealing with the larger datasets, Decision Trees perform well. As huge amount of data flows through the computer network, DTs are crucial. It is very much suitable in detecting the real time intrusions due to its high performance and it is helpful for Security officers to alter and examine these type of easily interpretable models created by DTs, which can also be used in rule based models with minimal processing. Advantageous property for ID model is the generalization accuracy of DTs.

After the implementation of Intrusion detection model, there will always be a possibility of occurring the little variations of known attacks i.e., new attacks on the system. Generalization accuracy of DTs will have the capability to detect new kinds of attacks.

The main focus is to separate the data into 2 classes i.e., "Attack" class and "Normal" class. Attack class includes the 4 kinds of attacks and they are DoS, R2L, Probe, and U2R. This process of separation is done for all the 5 classes. In order to separate the data into attack or normal type, training data is used to bulit a classifier and then with the implemented classifier the testing data is tested. For each of the 5 classes, accuracy is shown in terms of percentages and the time of training and the time of testing is shown in seconds.

V. EXPERIMENTAL RESULTS

Fig:6 and Fig:7 shows the accuracy, precision and recall results of SVM and ELM. Whereas Fig:8 and Fig:9 shows the accuracy, precision and recall measures of Random Forest and Decision Tree algorithm.

The SVM and ELM attained an accuracy of 0.9578 and 0.7105. The Random Forest and Decision Tree showed an accuracy of 0.8724 and 0.9618. SVM is better than RF and ELM. Hence, when comparing all the four algorithms, DT attained higher accuracy. We also measured precision and recall of all algorithms for detecting intrusions are shown in following figures.

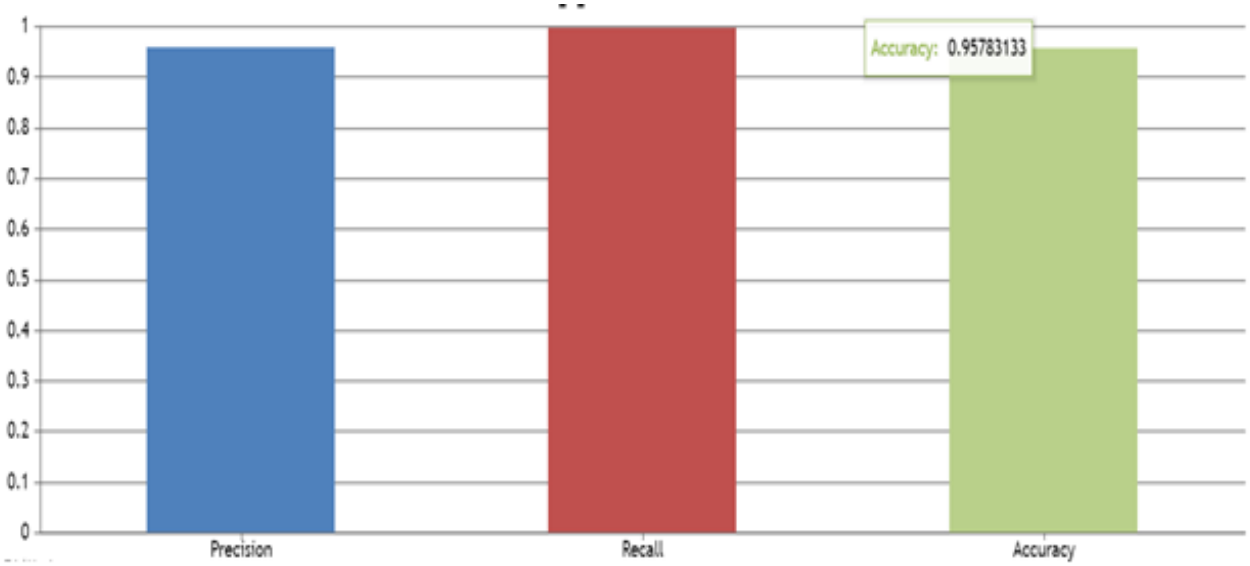

Fig. 6. SVM Results for intrusion detection.

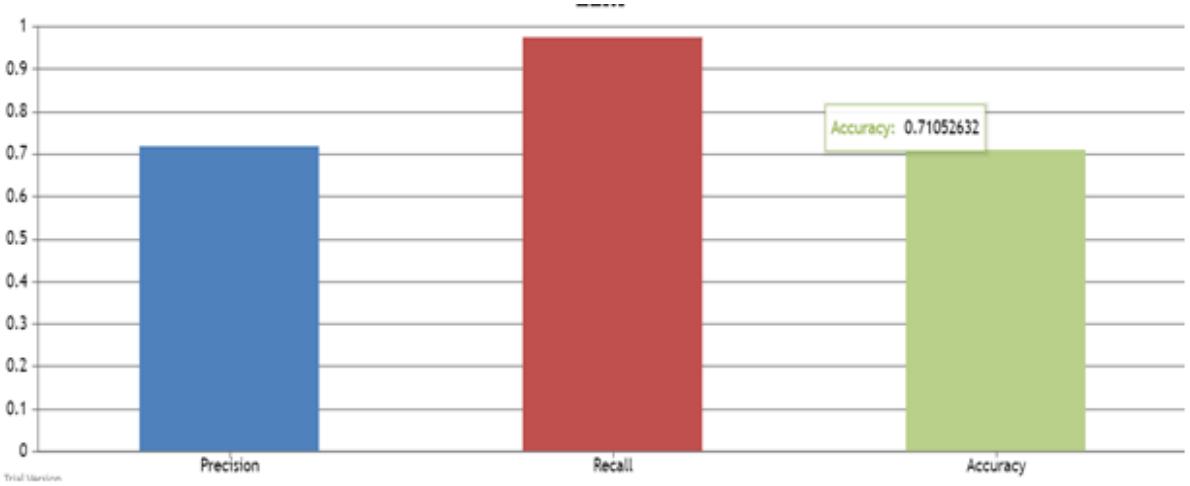

Fig. 7. ELM results for intrusion detection

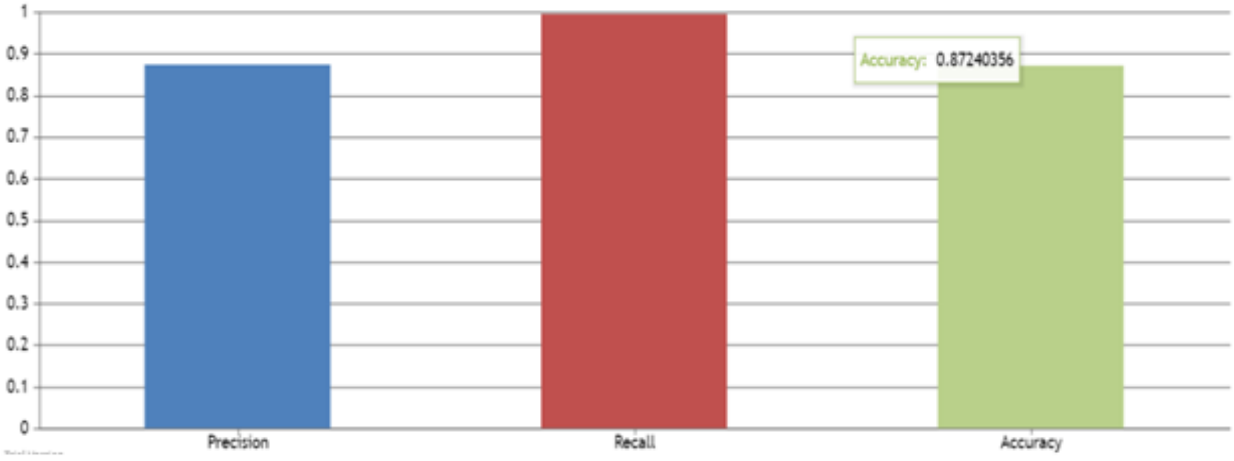

Fig. 8. Random Forest Results for intrusion detection.

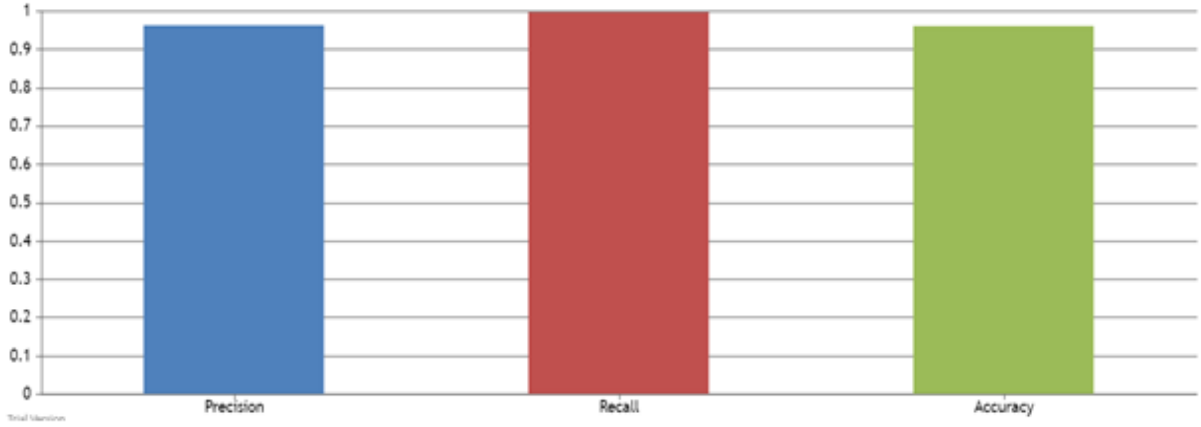

Fig. 9. Decision Tree Results for intrusion detection.

VI. CONCLUSION

As our day to day activities heavily rely on the forthcoming and present computer and network system, it is necessary to detect and prevent the intrusions. Besides, because of IOT, forthcoming challenges become more daunting. Due to this, from the last few decades IDS have become crucial. However, in the recent literature ML techniques have become common to solve this ID problem, eventhough many techniques and approaches have been used. Furthermore, some of the commonly used ML techniques are not suitable for examining the large data for detecting the intrusions of the system and network information.

In order to deal with this issue, 4 ML techniques are considered in this work and are compared with one another. They are RF, SVM, ELM and Decision Tree. On the complete/full data sample which containing both normal and intrusive activities, DT beats the other approaches in terms of precision and accuracy. Hence, a convenient and advisable technique for IDS for examining the larger data is DT.

REFERENCES

1. A. A. Olusola., A. S. Oladele and D. O. Abosede, “Analysis of KDD ’99 Intrusion Detection Dataset for Selection of Relevance Features,” Proceedings of the World Congress on Engineering and Computer Science I, San Francisco, 20-22 October 2010.
2. H. Altwaijry and S. Algarny, “Bayesian Based Intrusion Detection System,” Journal of King Saud University— Computer and Information Sciences, Vol. 24, No. 1, 2012, pp. 1-6
3. O. A. Adebayo, Z. Shi, Z. Shi and O. S. Adewale, “Network Anomalous Intrusion Detection using Fuzzy-Bayes,” IFIP International Federation for Information Processing, Vol. 228, 2007, pp. 525-530.
4. S. M. Bridges and R. B. Vaughn, “Fuzzy Data Mining and Genetic Algorithms Applied To Intrusion Detection,” Proceedings of the National Information Systems Security Conference (NISSC), Baltimore, October 2000, pp. 16-19.
5. Q. Wang and V. Megalooikonomou, “A Clustering Algorithm for Intrusion Detection,” Proceedings of the Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, Vol. 5812, 2005, pp. 31-38.
6. B. Pal and M. A. M. Hasan, “Neural Network & Genetic Algorithm Based Approach to Network Intrusion Detection & Comparative Analysis of Performance,” Proceedings of the the 15th International Conference on Computer and Information Technology, Chittagong, Bangladesh, 2012.
7. H.Wang, J. Gu, and S.Wang, ``An effective intrusion detection framework based on SVM with feature augmentation," Knowl.-Based Syst., vol. 136, pp. 130_139, Nov. 2017.
8. A. A. Aburomman and M. B. I. Reaz, ``A novel SVM-kNN-PSO ensemble method for intrusion detection system," Appl. Soft Comput., vol. 38, pp. 360_372, Jan. 2016.
9. F. Kuang,W. Xu, and S. Zhang, ``A novel hybrid KPCA and SVM with GA model for intrusion detection," Appl. Soft Comput., vol. 18, pp. 178_184, May 2014.
10. S. Teng, N.Wu, H. Zhu, L. Teng, andW. Zhang, ``SVM-DT-based adaptive and collaborative intrusion detection," IEEE/CAA J. Automatica Sinica, vol. 5, no. 1, pp. 108_118, Jan. 2018.
11. M. R. G. Raman, N. Somu, K. Kirthivasan, R. Liscano, and V. S. S. Sriram, ``An ef_cient intrusion detection system based on hypergraph_Genetic algorithm for parameter optimization and feature selection in support vector machine," Knowl.-Based Syst., vol. 134, pp. 1_12, Oct. 2017.
12. R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, ``A hybrid network intrusion detection framework based on random forests and weighted k-means," Ain Shams Eng. J., vol. 4, no. 4, pp. 753_762, 2013.
13. N. Farnaaz and M. A. Jabbar, ``Random forest modeling for network intrusion detection system," Proc. Comput. Sci., vol. 89, pp. 213_217, Jan. 2016.
14. A. Alazab, M. Hobbs, J.Abawajy, and M. Alazab, Using Feature Selection for Intrusion Detection System, International Symposium on Communications and Information Technologies, 2012.
15. C.-C. Chang and C.-J. Lin, ``LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1_27:27, 2011.
16. L. Breiman, “Random Forests,” Machine Learning, Vol. 45, No. 1, 2001, pp. 5-32.
17. G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, ``Extreme learning machine for regression and multiclass classification," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 2, pp. 513_529, Apr. 2012.

AUTHORS PROFILE

Katikela Haritha, is currently pursuing Master’s degree program in Computer Science and Engineering department in Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India. She has completed her Bachelor of Technology in Computer Science and Engineering department from G. Narayanamma Institute of Technology and Science in 2017, Hyderabad, Telangana, India.
E-mail: haritha.katikela@gmail.com

Dr. CH. Mallikarjuna Rao received the B. Tech degree in computer science from Dr. Baba Sahib Ambedkar Marathwada University, Aurangabad, Maharastra in 1998, M. Tech Degree in Computer Science and Engineering from JNTU Anantapur, Andhrapradesh in 2007 and Ph.D in Computer Science and Engineering from JNTU , Ananthapuramu. Currently he is working in “ Gokaraju Rangaraju Institute of Engineering and Technology”, Hyderabad. Telangana ,India. His area of Interests are Data Mining, Bigdata and Software Engineering.
E-mail: chmksharma@yahoo.com