DOI: https://doi.org/10.24113/ijoscience.v3i10.13

Analysis of Intrusion Detection and Classification Using Machine Learning Approaches

Anjum Khan
M.Tech. Scholar
Department of CSE
Sagar Institute of Research and Technology
Bhopal, M.P., India
anjumkhan347@gmail.com

Anjana Nigam
Professor
Department of CSE
Sagar Institute of Research and Technology
Bhopal, M.P., India
anjana.ngm@gmail.com

***Abstract*— As the network primarily based applications are growing quickly, the network security mechanisms need a lot of attention to enhance speed and preciseness. The ever eVolving new intrusion types cause a significant threat to network security. Though varied network security tools are developed, however the quick growth of intrusive activities continues to be a significant Issue. Intrusion detection systems (IDSs) are wont to detect intrusive activities on the network. Analysis showed that application of machine learning techniques in intrusion detection might reach high detection rate. Machine learning and classification algorithms facilitate to design "Intrusion Detection Models" which might classify the network traffic into intrusive or traditional traffic. This paper discusses some usually used machine learning techniques in Intrusion Detection System and conjointly reviews a number of the prevailing machine learning IDS proposed by researchers at different times. in this paper an experimental analysis is performed to demonstrate the performance analysis of some existing techniques in order that they will be used further in developing Hybrid Classifier for real data packets classification. The given result analysis shows that KNN, RF and SVM performs best for NSL-KDD dataset.**

Keywords— Intrusion Detection System, Anomaly Detection, Supervised learning, Unsupervised, Detection Rate.

I. INTRODUCTION

To detect intrusion or attack in the network or host computer one of the tool is Intrusion Detection System (IDS)s which can be either of host based or network based. A host based IDS detects attacks on a host computer whcih can be divided into four types, namely (a) File System Monitors, (b) Log file analyzers, (c) Connection analyzers, (d) Kernel-based IDS [1, 2]. Whereas, a Network Intrusion Detection System (NIDS) detects intrusions or attack into a network traffic which is generally installed at network gateway or server. Moreover, supported the data analyzing technique there are primarily two categories of IDSs, signature-based and anomaly based. A signature-based system detects attacks by analyzing network knowledge for attack signatures hold on in its information. This type of IDS detects antecedently known attacks, whose signatures are hold on in its database. On the other hand, an anomaly-based IDS appearance for deviations from traditional behavior of the subject. Anomaly-based systems are capable of detecting novel attacks [2]. Here some very common methods given which are used by intruders to gain control of computers are Trojan horse, Back door, Denial of Service, Email-borne Viruses, Packet sniffing, Spoofing, etc. It is clear from the study that a network packet has 42 features and the four simulated attacks such as:

Denial of Service (DoS) attack: Over usage of the bandwidth or non accessibility of the system resources ends up in the DoS attacks. Examples: Teardrop and Smurf.

User to Root (U2R) Attack: initially attacker access normal user account, later gain access to the basis by exploiting the vulnerabilities of the system. Examples: Perl, Load Module and Eject attacks.

Probe Attack: Have an access to entire network information before introducing an attack. Examples: ipsweep, nmap attacks.

Root to local (R2L) Attack: By exploiting a number of the vulnerabilities of the network offender gains native access by causing packets on a remote machine.

Machine learning techniques can be effective for detecting intrusions. Many Intrusion Detection Systems are modeled based on machine learning techniques. Learning algorithms are designed either on offline dataset or real data collected from university or organizational networks. Usually machine learning techniques is classified into 2 classes i.e. supervised Learning and unsupervised Learning. In supervised learning the training dataset is instantly accessible together with its target vector. The learner learns from available data taking guidance of the output vector. In contrast to supervised learning, unsupervised learning systems learn from its atmosphere. Systems learn from coaching knowledge; however there's no target vector accessible. Some usually used machine learning techniques within the field of

intrusion detection are like Artificial Neural Network (ANN), decision Tree, Support Vector Machine, Bayesian Classification, Self-organizing Map, etc.

II. RELATED WORK

Sufyan T. Faraj et al. [2] proposed BPANN based intrusion detection model for classification of abnormal network packets from normal traffic packets and accuracy of about 93% is achieved. Back-propagation Multi Layer Perceptron (MLP) based anomaly detection technique is used to identify normal users' profile was proposed by Ryan et al. [3]. Their MLP model evaluates the users' commands for possible intrusions at the end of each log session. The top 100 important commands used by the user throughout the session was used to determine the user's behavior. They used a 3 layer MLP model with two hidden layers and found that their MLP model was able to correctly identify 22 cases out of 24Similarly, a method primarily based intrusion detection approach that gives the flexibility to generalize from previously determined behavior to acknowledge future unseen behavior was proposed by Ghosh et al. [4]. Their framework employs artificial neural networks (ANNs) and may be used for each anomaly findion so as to find novel attacks and misuse detection so as to detect best-known attacks and their variations.

Meng et al. [8] analyzed ANN, SVM and DT plans for abnormality location in a uniform situation and reasoned that J48 calculation of DT gives better performance over the other two schemes. The detection rate of low successive attack analysis (U2R, R2L) was likewise high. Feng et al. [9] presented another classification technique and used the benefits of SVM and Clustering based on Self-Organized Ant Colony Network.

Sumaiya Thaseen Ikram et al. [9] proposed an intrusion detection system demonstrate utilizing chi-square feature selection and multi class support vector machine (SVM). A parameter tuning strategy is received for streamlining of Radial Basis Function portion parameter to be specific gamma spoke to by '!' and over fitting steady 'C'. These are the two critical parameters required for the SVM model. The principle thought behind this model is to build a multi class SVM which has not been received for IDS so far to diminish the preparation and testing time and increment the individual arrangement precision of the network attacks.

Manjula et al. [10] proposed an classification and predictive models for intrusion detection which is worked by utilizing machine learning order calculations to be specific Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. An experimental result shows that Random Forest Classifier out plays out alternate strategies in recognizing whether the information movement is normal or an attack.

Saad Mohamed et al. [11] presented a hybrid approach to anomaly detection using of K-means clustering and Sequential Minimal Optimization (SMO) classification.

Ibrahim et al. [12] in like manner connected a multi-level model with various machine learning procedures, for example, C5, MLP, and Naïve Bayes. The analysis used one of the techniques at each level to classify one category, thereby confirming that multilevel techniques exhibit higher detection accuracy than a single technique. To lessen the false alarm rate of irregularity based IDS, many machine learning procedures, including support vector machine (SVM) Feng et al. [13] connected extreme learning machine (ELM) alongside models consolidating a few methods. Each model offers particular qualities and shortcomings, with general nonexclusive identification rates relentlessly expanding. SVMs display great identification execution with IDSs as far as characterizing the stream of a system into normal or anomalous behaviors.

Horng et al. [14] proposed an IDS based on a combination of BIRCH hierarchical clustering using SVM technique. Their proposed method achieved a good accuracy of up to 95.72% with a false alarm rate of 0.7%.

Kuang et al. [15] proposed SVM with kernel principal component analysis (KPCA) and genetic algorithm (GA) based IDS. Dimension reduction using KPCA was used, whereas optimization is achieved using genetic algorithm. The average detection rate was 95.26%, whereas the average false alarm rate was 1.03%. ELMs exhibit performance comparable with that of SVMs in terms of classifying instances of IDS.

Gogoi, Bhattacharyya et al. [16] proposed a multi-level hybrid IDS using a combination of supervised, unsupervised, and outlier methods. This system was evaluated with three datasets, namely, real-time flow dataset, DDoS dataset, and the KDD Cup 1999 with NSL-KDD datasets. The system performance was good with a false alarm rate of 3.4% with the corrected KDD Cup 1999 dataset.

Wathiq Laftah Al-Yaseen et al. [17] proposes a multi-level hybrid intrusion detection model display that utilizations Support vector machine and outrageous learning machine to enhance the efficiency of recognizing known and unknown attacks. An modified K-means algorithm is additionally proposed to build a high quality training dataset that contributes altogether to enhancing the execution of classifiers. The popular KDD Cup 1999 dataset is used to evaluate the proposed model. Compared with other methods based on the same dataset, the proposed model shows high efficiency in attack detection, and its accuracy (95.75%) is the best performance thus far.

III. MACHINE LEARNING APPROACH

Artificial Neural Network *(ANN)s* are the computational models of neural structure of human mind. Neurons are the essential building squares of human mind. An ANN is a layered network of artificial neurons. An ANN may comprise of an input layer, at least one hidden layer(s) and a ouput layer. The artificial neurons of one layer are completely or mostly associated with the artificial neurons of the following layer. Each of these associations is related with a weight, and input associations with the past layers are additionally conceivable [2].

Decision tree is one of the least difficult machine learning methods. A decision tree can efficiently reflected as an arrangement of if-then guidelines. The arrangement begins from root node, navigating down the tree till the reasonable leaf node. Every node of the tree represents the solution. Every node tests on a property of the case and descending branch of that node corresponds to one of the values of that attribute. Beginning from the root node, every node tests the attribute determined by that node and moves down the tree through the branch coordinating the esteem, till it achieves a leaf node [10].

Support Vector Machine (SVM) maps the information vector into a higher dimensional element space. It is a binary classification method that orders input occurrences into two classes. Just the Support Vectors decide the ideal isolating hyper-plane to arrange input occurrence into one of the two classes. Support Vectors are the points closest to the separating hyper-plane. During classification, mapped input vectors set on one side of the isolating hyper-plane in the component space falls into one class and put on the opposite side of the plane falls into alternate class. In case the data points are not linearly separable, SVM uses suitable kernel function to map them into higher dimensional space, so that, in that higher dimensional space they become separable [9].

Bayesian learning is a statistical learning strategy in view of probabilities of hypotheses. An earlier probability is assigned to every hopeful hypotheses in light of prior learning. Training examples may increase or decrease the probability of a hypothesis to be correct. This likelihood can be computed utilizing Bayes' hypothesis. Classification is finished by joining predictions of multiple hypotheses, weighted by their probabilities. These probabilities in Bayesian strategy could be figured utilizing Bayes' hypothesis. Necessity of starting learning of numerous probabilities make practical application of Bayesian methods difficult [10].

Self-Organizing Map (SOM) is an exceptional class of unsupervised learning Artificial Neural Network. At first, every unit is assigned with weight vector. An input vector is contrasted and the weight vector of each unit of the SOM. The weights of the nearest unit and its neighbors are updated after every emphasis amid the preparation procedure. Once the preparation training is finished. Each input vector has a relating yield vector and the Euclidean separation between the input and every unit [10].

IV. PROPOSED MODEL

This section describes the proposed hybrid model for intrusion detection. The NSL-KDD dataset as well as real dataset can be used as a benchmark to evaluate the performance of the proposed model. The algorithm flow of the proposed method is described as follows:

Following steps will be used to build the proposed model for intrusion detection:

Step 1: Convert the symbolic attributes protocol, service, and flag to numerical.

Step 2: Normalize data to [0,1].

Step 3: Separate the instances of dataset into two categories: Normal, Attack.

Step 4: The data set is divided as training data and testing data.

Step 5: Train hybrid classifier with these new training datasets.

Step 6: Test hybrid model with dataset.

Step 7: Finally computing and comparing TPR, FPR, Precision, Recall, F1-Score and Accuracy for different classifier or IDS models.

The proposed algorithm flow diagram of intrusion detection model is illustrated in figure 1. The proposed framework consists of three phases i.e. Preprocessing, Post Processing Phase and Intrusion Detection Phase. Below each stage is described individually in details.

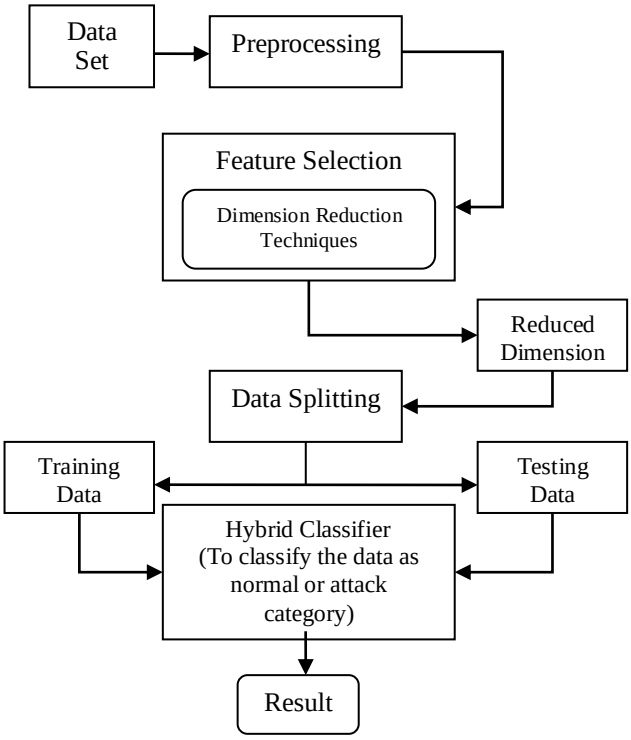


Figure 1: Proposed Flow Diagram of Intrusion Detection System

V. IDS TERMINOLOGIES

To evaluate the proposed algorithm, it is concentrated on three indications of performance: detection rate, accuracy and False Positive Rate (FPR).

If one sample is an anomaly and the predicted label also stands anomaly, then it is called as true positive (TP).

If one sample is an anomaly, but the predicted label stands normal, then it is called as false negative (FN).

If one sample is a normal and the predicted label also stands normal, then it is true negative (TN).

If one sample is normal, but the predicted label stands anomaly, then it is termed as false positive (FP).

TP stands the number of true positive samples, FN stands the number of false negative samples, FP stands the number of false positive samples, and TN stands the number of true negatives.

From equation (i) to (iii), the detection rate, accuracy, False Positive rate (FPR) is achieved.

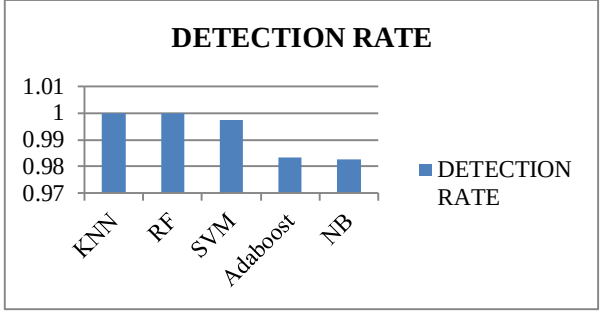
Detection Rate = TP/(TP+FN) (i)

Accuracy = (TP+TN)/(TP+TN+FP+FN) (ii)

FPR= FP/(FP+TN) (iii)

VI. RESULT ANALYSIS

An Experimental analysis is performed using proposed architecture to demonstrate the performance evaluation of some existing techniques because they can be used further in developing Hybrid Classifier. Tabular summarization of the experimentally reviewed models is also presented in Table I.



The experimental analysis is performed on NSL-KDD dataset having 30 features using 20% of the training dataset. The performance evaluation is done by using some existing classifiers such as Support Vector Machine (SVM), Random Forest (RF), Adaboost, K Nearest Neighbour and Naïve Bayes. Table I shows the performance evaluation of different techniques with respect to Detection Rate (DR), Accuracy and False Positive Rate (FPR) and their respective graphs are shown if Figure 1-3.

Table I: Comparative Analysis of Classifiers

Techniques	DR	Acc	FPR
KNN	0.9999	1	0.00008267
RF	0.9998	0.9999	0.00016534
SVM	0.9973	0.9962	0.0023
Adaboost	0.9835	0.9851	0.0143
NB	0.9828	0.9453	0.0136

Figure 1: Comparative analysis of DR

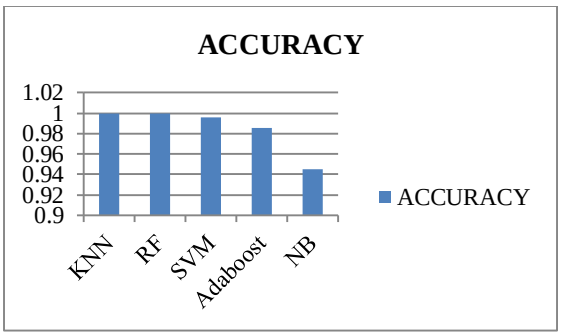


Figure 2: Comparative analysis of Accuracy

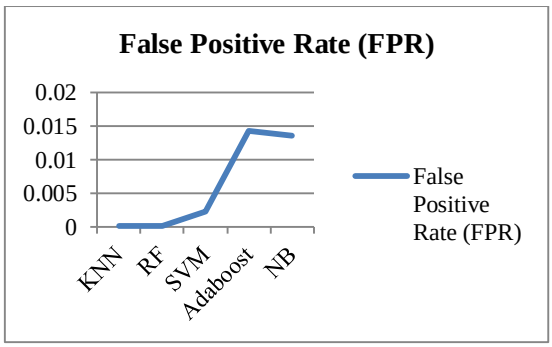


Figure 3: Comparative analysis of FPR

VII. CONCLUSION

In modern society, the security of computer networks becomes an increasingly vital Issue to be solved. Traditional intrusion detection techniques lack extensibility in face of changing network as well as adaptability in face of unknown attack type. Machine learning techniques are proved to be efficient for intrusion detection. High accuracy in intrusion detection can be achieved using machine learning techniques even though the detection accuracy depends on some other factors too. Some of them are selection of correct feature set, selection of appropriate training and testing data, etc. With the selection of the appropriate attributes for these factors, a higher performance could be achieved. In this paper an experimental analysis is performed using proposed architecture to demonstrate the performance evaluation of some existing techniques because they can be used further in developing Hybrid Classifier for real data packets classification. The given result analysis shows that KNN, RF and SVM performs best for NSL-KDD dataset. This result analysis can be used as reference in future for developing an intrusion detection system for real data packets.

REFERENCES

[1] 2. Garcia-Teodoro, P., "Anomaly-Based network intrusion detection: techniques", systems and challenges. Comput. Security Vol. 28.Issue, pp. 18–28, 2009.

[2] Sufyan T Faraj Al-Janabi, Hadeel Amjed Saeed, "A neural network based anomaly intrusion detection system", IEEE, 2011.

[3] J. Ryan, M. Lin, and R. Miikkulainen, "Intrusion Detection with Neural Networks," Conference in Neural Information Processing Systems, 943–949.

[4] A. K. Ghosh and A. Schwartzbard, "A Study in Using Neural Networks for Anomaly and Misuse Detection," Conference on USENIX Security Symposium, Volume 8, pp. 12–12, 1999.

[5] P. L. Nur, A. N. Zincir-heywood, and M. I. Heywood, "Host-Based Intrusion Detection Using Self-Organizing Maps," in Proceedings of the IEEE International Joint Conference on Neural Networks, pp. 1714–1719, 2002.

[6] K. Labib and R. Vemuri, "NSOM: A Real-Time Network-Based Intrusion Detection System Using Self-Organizing Maps," 2000.

[7] Sharma, R.K., Kalita, H.K., Issac, B., "Different firewall techniques: a survey", International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2014.

[8] Meng, Y.-X., "The practice on using machine learning for network anomaly intrusion detection", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 2, IEEE, 2011.

[9] Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", Journal of King Saud University –Computer and Information Sciences, 2016.

[10] Manjula C. Belavagi and Balachandra Muniyal, "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Computer Science", Elsevier, 2016.

[11] Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, "Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique", International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.

[12] Ibrahim, H. E., Badr, S. M., & Shaheen, M. A., "Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems", International Journal of Computer Applications, Vol. 56, Issue 7, pp. 10–16, 2012.

[13] Wen Feng, Qinglei Zhang, Gongzhu Hu, Jimmy Xiang Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks", Elsevier, Vol 37, pp 127-140, 2014.

[14] Shi-JinnHorng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines" Expert Systems with Applications, Elsevier, Vol. 38, pp. 306–313, 2011.

[15] Kuang, F., Xu, W., & Zhang, S., "A novel hybrid KPCA and SVM with GA model for intrusion detection", Applied Soft Computing Journal, Vol. 18, pp. 178–184, 2014.

[16] Prasanta Gogoi, D.K. Bhattacharyya, B. Borah1 and Juga, K. Kalita, "MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method", The Computer Journal, Vol. 57 Issue 4, pp. 602-623, 2014.

[17] Wathiq Laftah Al-Yaseen , Zulaiha Ali Othman ,Mohd Zakree Ahmad Nazri, "Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System", International Journal in Expert Systems With Applications, Elsevier, 2017.

[18] He, L., "An improved intrusion detection based on neural network and fuzzy algorithm. Journal of Networks, Vol. 9, Issue 5, pp. 1274–1280, 2014.

[19] Hoque, M. S., Mukit, M. A. , & Bikas, M. A. N., "An implementation of intrusion detection system using genetic algorithm", International Journal of Network Security & Its Applications, Vol 4, Issue 2, pp. 109–120, 2012.