Network Traffic Classification Model Using KNN Classifier

Aditi Sharma
aditisharmabknr@gmail.com
Govt. engineering college of bikaner and technology
Bikaner

Dr. Subhash panwar
panwar.subhash@gmail.com
Govt. engineering college of bikaner and technology
Bikaner

ABSTRACT - An intrusion is the activity through which such an event is triggered that compromises the security of overall system. Within the system, either, internal or external intrusions arise. It can also be known as the illegal action due to which the computer becomes hazardous. The network intrusion technique has various phases which include like pre-processing, feature extraction and classification. The KNN classification model is modified to improve accuracy, precision and recall values. The proposed model is implemented in python and it is compared with the existing models like SVM, ensemble for the intrusion detection. It is analyzed that proposed model give high accuracy, precision and recall as compared to existing models.

KEYWORDS - IDS, Machine learning, SVM,KNN

I. INTRODUCTION

An intrusion is the activity through which such an event is triggered that compromises the security of overall system. Within the system, either, internal or external intrusions arise. It can also be known as the illegal action due to which the computer becomes hazardous [1]. The system that is designed to monitor the events performed in a computer network and can identify the security related problems of networks easily is known as an intrusion detection system. It acts as an alarm system through which any kinds of violations can be alarmed to the authorities. The systems can also be alerted about any kinds of false messages or mails. It acts as a tool through which the system is secured against any kinds of intrusions or attacks. The most important objectives of IDS are checking the attack scenarios and providing the support for defense management. IDSs are being used by all the applications including networking in them to provided security. The malicious activities which are impossible to be detected by even firewalls can be detected through IDS. The attacks are possible on computer systems since there are highly sensitive services and computer applications involved. Due to the intrusions, the computer applications face data driven attacks, since these attacks are caused in sensitive services and there are unauthorized logins in sensitive files. There are certain principles included in Intrusion Detection based on the static and dynamic attack patterns among which few are explained below:

A. Anomaly Detection: The network traffic is monitored continuously in order to detect the anomalies such that an activity can be detected as being either normal or malicious. The intrusion activity is determined through regular observation and comparison of network traffic with baseline [5]. No prior information is required by this attack for detecting intrusions. However, a few demerits of this system are fault diagnosis and high false positive rates.

B. Misuse Detection: The signature-based detection is the system through which the patterns or signatures of possibility of an attack are identified. High accuracy is provided through this detection method and for the recognized attacks, the variations are detected efficiently using this method. However, it is not possible to detect new kinds of attacks since the patterns and signatures are absent. A hybrid method is designed if the misused IDS and anomaly are merged together.

Intrusion detection system is known as the system using which any kinds of attacks possible to cause intrusions in the working of systems are identified. The applied system is commonly known as IDS. The intrusions are also commonly known as attacks or anomalies [6]. This approach helps in examining the activities of a network or system. The IDS is categorized based on detecting intrusion methods.

a. Host-Based IDS: There are the systems that monitor the devices on which they have been deployed. For the execution of monitoring program, this method monitors the states of main system using the audit logs for program execution. Due to the dependency of HIDS on audit logs, they are limited. Another issue faced here is the sheer volume of audit logs.

Every monitored log requirement needs to be parsed here. Therefore, the performance of host system will be affected if HIDS is installed. Because of the vulnerability in audit files, the integrity of HIDS is affected directly [7]. This is known to be another demerit of this system. Because of the changes in audit file, seeing and detecting what happened is not possible in HIDS.

II. LITERATURE REVIEW

Pratibha Khandait, et.al (2020) analyzed that the DPI was very significant technique for traffic classification [8]. The Deep Packet Inspection was very high cost operation as the string matching amid payload and application signature was performed in it. The traffic classification methods that had utilized earlier categorized the flows of application by performing a number of scans of payload. The words were extracted in first scan. The matching of words with application signatures was done in second scan. An approach was suggested in this paper for the classification of flow of network with single scan of payloads. The heuristic technique was carried out so as a sub-linear search complexity was obtained. Some payload's initial bytes scanned in it. The potential application signature had verified to match subsequent signature. A large dataset was employed to conduct the experiment. There were 171873 network flows included in this dataset. The experimental results demonstrated that the classification precision achieved from it was evaluated 98%.

Ibraheem Saleh, et.al (2020) suggested a Network Traffic Images that included 2D formulation of a packet header length stream for network traffic classification [9]. The deep convolutional neural networks had utilized in this technique. The suitable scheme was assumed so as the 1D packet sub-flow was transformed into a two-dimensional. For this purpose, the designing of various five network traffic image orientated mappings were completed. The packet-relative and the time-relative were two mapping strategies that had carried out. The packet flow's packets were mapped to the pixels in image. The experiments demonstrated that the network traffic images were utilized in deep learning and the highest classification precision had achieved.

Auwal Sani Iliyasu, et.al (2020) recommended a semi-supervised learning approach to classify the network traffic in which DCGAN was carried out [10]. The generators of Deep Convolution Generative Adversarial network produced the samples. The performance of a classifier was enhanced using these samples and unlabeled data. The difficulties had alleviated that was associated with the large dataset. A self-collected dataset of QUIC protocol and dataset ISCX VPN-Non-VPN had carried out for the evaluation of this model. The efficiency of this approach was proved. The accuracy acquired from the QUIC was counted as 89% and the accuracy acquired from other dataset was evaluated 78% with a smaller number of labeled samples.

Dongpu Li, et.al (2020) suggested a new framework in which unsupervised domain adaptation was utilized for classifying cross-domain network traffic [11]. It was the initial attempt to deal with this issue. The marginal and the conditional distribution of training data as well as unlabelled data had adapted using feature transformation module with current task. The classifier was re-trained with the transformed data. The varied network conditions altered the statistical attributes that results in increasing the accuracy for classifying the traffic. Five commonly used models were implemented for the evaluation of suggested framework. The classification accuracy obtained in the experimental outcome was computed 86%.

Edyta Biernacka, et.al (2020) focused on the solution of dynamic routing problem with traffic classification in which Software Defined Elastic Optical Networks framework was utilized [12]. The 3 types of traffic had described that were classified on the basis of CISCO Traffic reports. Two algorithms were suggested for dealing with the Routing, Modulation and Spectrum Assignment problem. For this purpose, the Split Spectrum and the Buffer algorithms were carried out. The representative network topologies and multiple simulations had employed to evaluate this approach. The efficient performance had achieved from Butter algorithm in the results. It was also demonstrated that the request blocking was diminished to acquire the highest priority traffic.

FaizZaki, et.al (2020) studied that the whole network traffic classification process was analyzed and emphasized on the classification methods in previous studies [13]. The exposure for classification granularity was also required in these studies as its application were increased in modern networks. The optimization of traffic classification was facilitated after comprehending the different levels and use cases classification granularity. The various levels and use cases of classification

granularity had investigated that was the main purpose of this paper. Consequently, the classification granularity was assembled into a systematic multilevel taxonomy that supported to obtain the deeper understanding of their applications. At last, the challenges and future directions were defined in this work.

FakhroddinNoorbehbahani, et.al (2018) presented a novel semi supervised algorithm for network traffic classification [14]. The recommended algorithm was derived from x-means clustering algorithm and a latest label propagation method. The assessment outcomes revealed that the label propagation method achieved classification accuracy of 95% which was more than average. The recommended label propagation technique performed the training of two classifiers called J48 and Naïve Bayes. Both of these classifiers achieved accuracy rate somewhat similar or sometimes above the accuracy rate achieved with real completely labeled dataset. The recommended approach could be improved in nearby future by applying a semi-supervised feature selection technique. In addition, the testing of recommended approach could be done on fresh network traffic datasets. An advanced adaptation of label propagation approach could be designed in future.

Jaehwa Park, et.al (2013) presented an analysis of network status classification in several operational networks with different scales [15]. This work was based on the two main concepts. According to the first concept, network traffic repeated cycles of congested states. The second concept was that the deviation of network latency was robustly connected with the background of the latency. This work repeatedly measured peer to peer network latencies in genuine operational networks and classified the status of network traffic regarding the stability and the burstiness of the latencies. The tested outcomes revealed that the recommended approach could evaluatethe status of network traffic and reflected the correlated variations.

Jing Ran, et.al (2018) implemented various sizes of CNN (Convolutional Neural Network)for the classification of network traffic [16]. In this work, the traffic flow had been processed as time sequence, images and videos correspondingly. The implementation of 3D CNN has been carried out for the first time in this work for classifying network traffic. The tested outcomes revealed that the use of both spatial and temporal attributes obtained maximum accuracy. It was concluded that the deep learning could give good performance in the classification of network traffic. The future work would be focused on learning the mixture of feature extractor and conventional classification models.

Yu Wang, et.al (2012) recommended a new technique for network traffic classification. This technique made use of a labeled training dataset as input and generated a set of signatures to match the application classes conferred within the data [17]. There were mainly four steps included in the recommended technique. These steps were known as pre-processing, tokenization, multiple sequence alignment and signature construction. In the first step, the extraction of application session payload was carried out. The second step discovered general substrings and integrated location limitations. The aim of third step was to convert the outcomes into usual lexis. This work made use of a real-time complete payload traffic trace for evaluating the recommended framework. In this framework, signatures for various applications were determined in automatic manner. The achieved outcomes revealed that the signatures were of extremely good quality and exhibited low FNs (False Negatives) and FPs (False Positives).

III. RESEARCH METHODOLOGY

This work is focused on to implement NTC (Network Traffic Classification) approach for classifying network as malevolent or non-malevolent. The malevolent behaviors of active users can be predicted using this approach. The recommended technique classifies network traffic in three steps. At first, the implementation of K-means clustering algorithm is done to cluster data on the basis of similarity and non-similarity. In order to refine the given dataset as input, it is imperative to eliminate several issues from it. These issues include data redundancy, missing values etc. The center point of network is computed using K-means clustering algorithm. Here, the computation of arithmetic mean of the entire dataset is carried out. Euclidian distance is measured from the center point for distinguishing like and unlike data objects. The similar data objects are placed in single cluster while the different clusters contain unalike data objects. Finally, the classification of data objects is carried out among two different classes using SVM classifier. This work makes use of KNN classifier for the clustering of non-clustered data objects. This phenomenon improves the accuracy and efficiency of classification algorithm. This algorithm measures Euclidian distance and differentiates alike and unalike data classes.

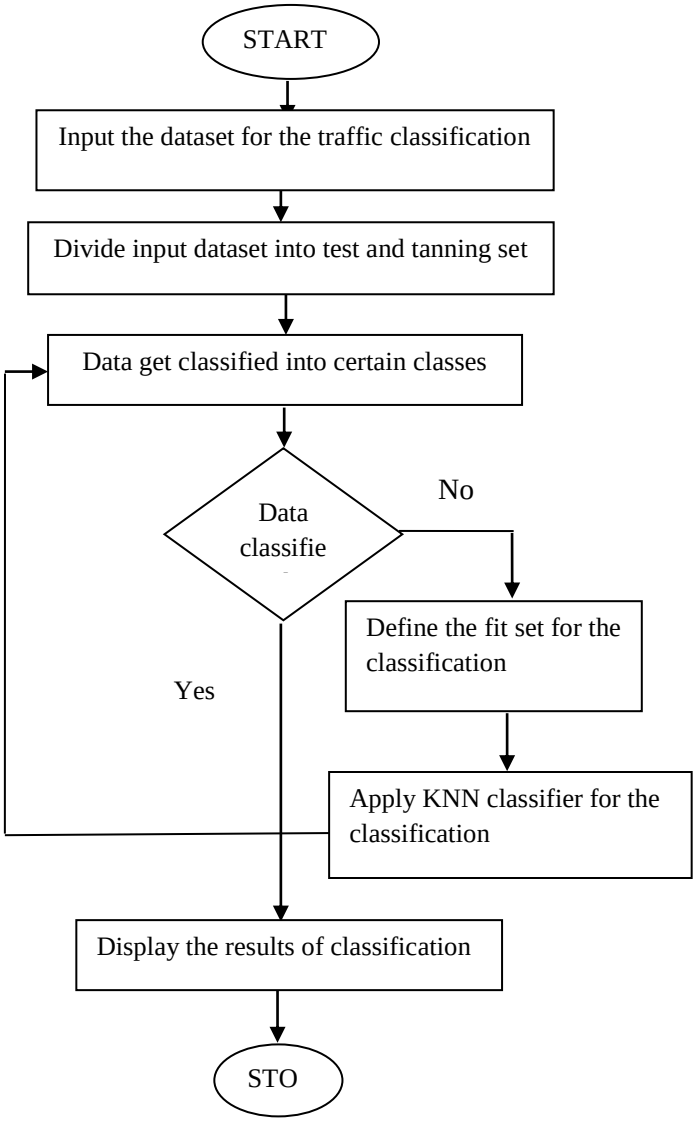

Figure 1: Flowchart of Methodology

IV. RESULT AND DISCUSSION

The first main scarcity in KDD data set is the occurrence of huge volume of unnecessary records. As per the analysis of KDD training and test sets, the training and testing set contain approx 78% and 75% of duplicate records respectively. The learning algorithm may be partial towards the most common records due to the availability of large numbers of unneeded records within the training set. Hence, it is necessary to forbid the infrequent and risky records. These records can cause harm to the networks. The techniques having improved detection rates on regular records provide support for generating biased results since repeated records occur in the test set. Also, this work makes use of twenty-one learned machines along with analyzing complexity level of records in KDD data set to label the records of entire KDD training and test sets. Hence, twenty-one predicted labels are offered for every record. As per the observation, the accurate classification of approx. 98% and 86% of records in the training and test set is carried out with all the twenty-one learners.

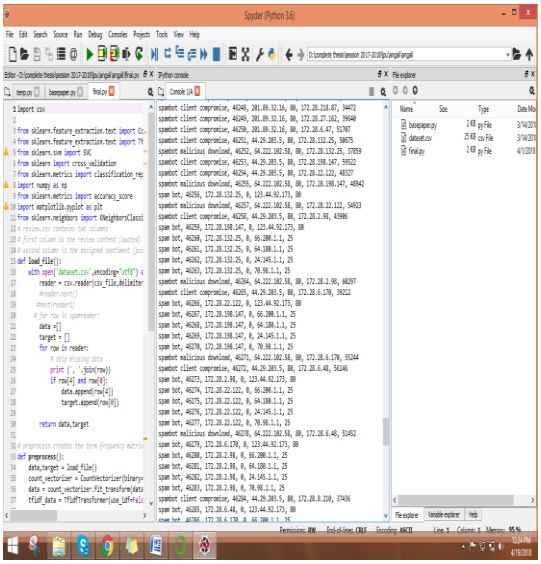

Figure 2: Apply KNN classifier

Figure 2 shows the classification of data into several classes by KNN classifier. The classification of classes is done along with compromised server namely DDoS. This classifier achieves classification accuracy of approx. 84%.

Accuracy: Accuracy is defined as the number of points correctly classified divided by total number of points multiplied by 100, as shown in eqn. 5.1.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

Precision: In pattern recognition, information retrieval and binary classification, precision (also called positive predictive

value) is the fraction of relevant instances among the retrieved instances.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

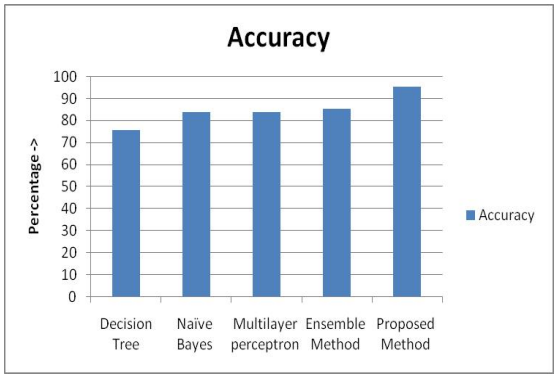

Figure 3: Accuracy Analysis

The figure 3 illustrates that a variety of models including DT, NB, multilayer perceptron, ensemble and proposed models are compared concerning accuracy. The analytic results reveal that the proposed model achives highest accurcay rate of almost 95% by performing better than other classifiers for predicting heart disorders.

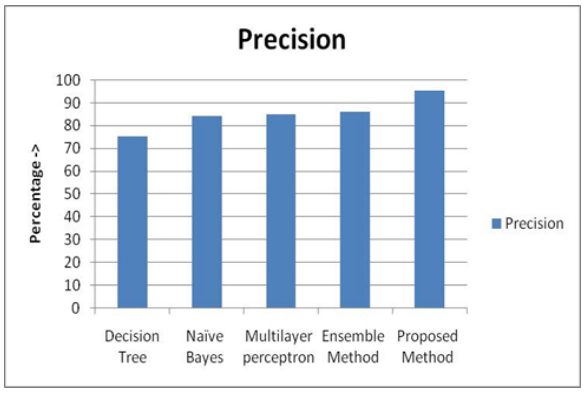

Figure 4: Precision analysis

As shown in figure 4, the various models of including DT, NB, MLP, ensemble and proposed models are compared in terms of precision. The analytic results reveal that the proposed model achives highest precison rate of almost 95% by performing better than other classifiers for predicting heart disorders.

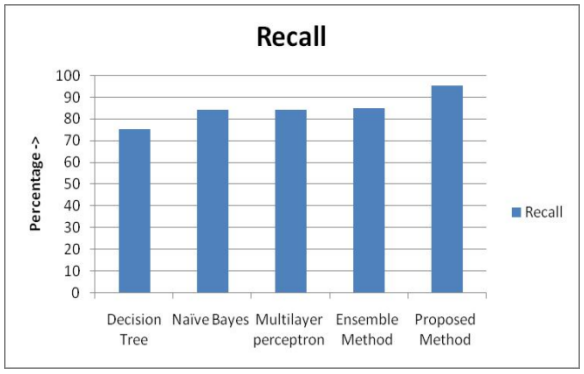

Figure 5: Recall Analysis

As shown in figure 5, the various models like DT, NB, multilayer perceptron, ensemble are compared with the new model in terms of recall. It is analyzed that recall of proposed model for heart disease prediction is approx 95 percent which is higher than the other models

V. CONCLUSION

Data classification is a major step in ML (Machine Learning). The developed computer programs perform classification for getting labeled datasets. The comparison of predicted heart rate and actual heart rate is carried out to determine whether the heart rate of patients is normal or not. Machine learning algorithms can be divided into two categories i.e. supervised learning and unsupervised learning. The tuning of appropriate parameters plays an important role in these classifiers. In this research work, the KNN classifier is proposed to increase performance for network traffic classification. The performance of proposed model is analyzed in terms of accuracy, precision and recall. It is analyzed that proposed model improve accuracy, precision and recall as compared to existing models for network traffic classification.

VI. REFERENCES

[1] Amrita, Kiran Kumar Ravulakollu, "A Hybrid Intrusion Detection System: Integrating Hybrid Feature Selection

Approach with Heterogeneous Ensemble of Intelligent Classifiers", International Journal of Network Security, Vol.20, No.1, PP.41-55, Jan. 2018

[2] Bayu Adhi Tama and Kyung-Hyune Rhee, "Performance evaluation of intrusion detection system using classifier ensembles", Int. J. Internet Protocol Technology, Vol. 10, No. 1, 2017

[3] M. Paz Sesmero, Agapito I. Ledezma and Araceli Sanchis, "Generating ensembles of heterogeneous classifiers using Stacked Generalization", WIREs Data Mining KnowlDiscov 2015, 5:21–34

[4] Necati DEMIR, Gokhan DALKILIC, "Modified stacking ensemble approach to detect network intrusion", 2018, Turkish Journal of Electrical Engineering & Computer Sciences, 26: 418-433

[5] Nanak Chand, Preeti Mishra, C. Rama Krishna, Emmanuel Shubhakar Pilli and Mahesh Chandra Govil, "A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection", 2016, IEEE

[6] M. Mazhar, U. Rathore, "Threshold-based generic scheme for encrypted and tunneled Voice Flows Detection over IP Networks", Journal of King Saud University Computer and Information Sciences, vol. 27, pp. 305–314, 2015.

[7] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, F oudilAbdessamia, "Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms", 2016 2nd IEEE International Conference on Computer and Communications, vol. 8, pp. 2451-2455, 2016.

[8] Pratibha Khandait, NeminathHubballi, Bodhisatwa Mazumdar, "Efficient Keyword Matching for Deep Packet Inspection based Network Traffic Classification", 2020, International Conference on COMmunication Systems &NETworkS (COMSNETS)

[9] Ibraheem Saleh, Hao Ji, "Network Traffic Images: A Deep Learning Approach to the Challenge of Internet Traffic Classification", 2020, 10th Annual Computing and Communication Workshop and Conference (CCWC)

[10] Auwal Sani Iliyasu, Huifang Deng, "Semi-Supervised Encrypted Traffic Classification with Deep Convolutional Generative Adversarial Networks", 2020, IEEE Access, Volume: 8

[11] Dongpu Li, Qifeng Yuan, Tan Li, Shuangwu Chen, Jian Yang, "Cross-domain Network Traffic Classification Using Unsupervised Domain Adaptation", 2020, International Conference on Information Networking (ICOIN)

[12] Edyta Biernacka, Michal Aibin, "On advantages of data driven traffic classification for dynamic routing in optical networks", 2020, International Conference on Computing, Networking and Communications (ICNC)

[13] FaizZaki, Abdullah Gani, Nor BadrulAnuar, "Applications and use Cases of Multilevel Granularity for Network Traffic Classification", 2020, 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)

[14] FakhroddinNoorbehbahani, SadeqMansoori, "A New Semi-Supervised Method for Network Traffic Classification Based on X-Means Clustering and Label Propagation", 2018, 8th International Conference on Computer and Knowledge Engineering (ICCKE)

[15] Jaehwa Park, JunSeong Kim, "A classification of network traffic status for various scales networks", 2013, The International Conference on Information Networking (ICOIN)

[16] Jing Ran, Yexin Chen, Shulan Li, "THREE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK BASED TRAFFIC CLASSIFICATION FOR WIRELESS COMMUNICATIONS", 2018, IEEE Global Conference on Signal and Information Processing (GlobalSIP)

[17] Yu Wang, Yang Xiang, Wanlei Zhou, Shunzheng Yu, "Generating regular expression signatures for network traffic classification in trusted network management", Journal of Network and Computer Applications, Volume 35, Issue 3, May 2012, Pages 992-1000