Machine learning Approach for the Network Traffic Classification

Riya Sharma[1], Ms. Shivani[2]
[1]M.Tech Student, [2]Assistant Professor
Rayat Institute of Engineering and Technology, Railmajra

Abstract - In the networked data, certain interconnected entities need to generate inferences. For example, for the interconnected web pages hyperlinks are available, the research papers have references and the conceivable terrorists can be linked through phone calls and accounts available in traced conversations. It is very common to find the ubiquitous nature of networks. There are social, financial, physical and transaction networks included in everyday lives of users. This research studies the models that define the influence of sensor nodes that are deployed in networks. Huge focus is presented on the models which help in predicting an important attribute based on the observed attributes. SVM classifier is applied on the collected data such that the data can be categorized as malicious or non-malicious.

Keywords: Machine Learning, Network traffic classification, SVM, KNN

I. INTRODUCTION

When two or more computers are connected with each other with the aim of achieving particular benefit, a network is formed. Based on the requirements of users involved in these networks, the information is forwarded or exchanged across the systems. The computer devices which can perform communication among each other are included in this set up [1]. To transmit the information and establish connections several resources are deployed in these networks. Information is provided by the network technology included in them. Many hardware and software resources are included when the devices are shared. A protocol that uses internet or other data networks to perform communication is called Voice over Internet Protocol (VoIP) [2]. Traditionally the Public Switched Telephone Network (PSTN) was used which however, had some efficiency and quality related issues. Internet is being used very commonly in voice based communications such that the cost of devices, operations and maintenance can be reduced. Earlier the telephone lines were used for VoIP which are now replaced by IP protocols such that the people can communicate with each other through voice. To assign the technology in a niche market, few factors are responsible [3]. They include the property standards, huge price tag as well as limited integration. The network scenarios have changed since the cost of open source VoIP tools is less and chances of risks are also reduced [4]. Using VoIP technology, clients are provided with additional standard xDSL at the least possible cost. Through the advancements in VoIP, it is possible to include the convergent networks that support voice and video services since in the traditional PSTN services this was not possible. However, to avoid detection or tax payments, few of the telecom operations charge intentionally even when these services are completely free of cost [5]. The goal of such providers is to get rid of the taxes that are charged by the government. It any illegal telecom traffic is performed, several threats are faced by the national security. Therefore, losses are suffered by the existing operators and national exchequers. To hide the identity, an illegal telecom gray traffic is available that brings all the calls external to the country and considers them as local calls [6]. Data analysis is a method that helps in collecting the data from different sources and performing evaluations and reviews on it. This method is of huge importance since elimination of redundant information is done and the data cleaning methods provide accurate conclusions. Thus, the data is cleaned, inspected and transformed through this method by applying different tools and techniques [7]. The major objective of data analysis is to recognize the information through which decision-making process is being supported. For data analysis, several methods are being included which include visualization, business intelligence and data mining. The results summarization and data analysis are performed to examine and interpret the important information. This method also helps in determining the data's quality and researchers can be provided with answers to the questions. To provide solution to any problem such that particular quality based results can be achieved, various statistical technologies have been applied [8]. When such techniques are applied, relationships among different variables are drawn. Parametric and non-parametric test are the two categories of statistical techniques. The different kinds of threats or attacks faced in VoIP and their effects on the overall systems are studied in this section. The networks face Denial of service (DoS) attack when no resources are available for the user. The server is filled with fake requests by the attacker such that no genuine requests are completed. For example, a server can only handle 100 users at a time. Now, the server is exhausted when it replied so many messages [9]. Therefore, the services of legal user are not fulfilled since the resources are being used completely by the requests of attacks. The attacker aims of

overhear the private data that is being exchanged among two parties. The attacker positions itself in between the two communicating vehicles to cause this attack. The attacker now performs communication among the sender and receiver [10]. However, a genuine user assumes that the communication is being held among users which are genuine only. The attacker overhears the communication among vehicles. Also, in a communication the modified message is added.

II. LITERATURE REVIEW

Mario A. Ramirez-Reyna, et.al (2017) designed a new mechanism for VoIP networks which was named as call admission control (CAC). The different codecs and/or codec mode-sets were applied and significantly investigated by this approach [11]. A factor whose value was based on the data rate of transferring obligation of that codec and/or codec mode-set was used in this method. To evaluate the performance of proposed approach, the packet level scrutiny and combined link were used. It was seen that for varying data rate transferring obligation ratios and proportion of clients, the proposed method attained the highest Erlang capability. Thus, the overall performance of systems was improved as per the mathematical outputs generated towards the end.

Murizah Kassim, et.al (2017) studied that from Third Generation (3G) to Fourth Generation (4G) networks, the wireless mobile telecommunication was designed. The initial two platforms included here were Mobile WiMAX and Long Term Evolution. Here, conference calls or voice operations could be performed by associating this method. To evaluate the performance of VoIP networks, a comparative analysis was performed on 3G and 4G networks [12]. It was possible to perform testing on voice Skype applications and collect the information to perform further tests. For analyzing the traffic and showing the level of performances, Jperf software was utilized. The performance of jitter, latency, bandwidth accessibility and latest LTE4G analysis were studied to test the overall performance results. It was seen that for both 3G and 4G LTE networks, average of 4 rating was provided. To provide good connections for both the links, it was important to include fundamentals like bandwidth, latency and jitter.

Jan Holu, et.al (2018) performed an analysis of around 16 million live calls collected over the IP-based telecommunication networks. Inspecting the dependence among the standard call period and call quality as required by the client was the major objective of this study. Non-monotonic link during duration and attributes was suggested unexpectedly by this analysis [13]. The universal assumption that longer calls were resulted in case of high quality was contradicted here. It is important to reconsider the usage of standard call duration as a displayer using the new discovery. The modeling of consumer performance was also affected by the outputs. The user behavior was not considered to be completely natural based on the discovery. Based on some peripheral aspects and performances of network, the resulting behavior was also influenced.

Eko Ramadhan, et.al (2017) studied that as a mode of interaction among two devices, the computer network technology gained huge growth. With respect to the communication media, the growth was increased [14]. Currently, to perform communication across networks, VoIP method was used. So far, one of the fastest emerging internet applications was VoIP. Using GNS3 adversary, the Asterisk applications were performed in the form of a server such that Private Automatic Branch eXchange (PABX) could be achieved. This research aimed to achieve optimum QoS value using diverse bandwidth by using a routing protocol named BGP. Based on the benchmark ITU-T G.114, the proposed method provided better results in terms of various performance parameters.

Mohammad Tariq Meeran, et.al (2017) proposed a research which aimed to improve the service quality of VoIP by designing two new methods. At first, to integrate the standards, protocols and voice codecs, the most appropriate selections were suggested. Also, for the mesh topology, the motionless and mobile supportive mesh nodes were added secondly [15]. The packets forwarding process was aimed to be designed here. With the scale of 0.2 in no mobility, 2.2 in partial mobility and 0.9 in full mobility situations, the quality of VoIP could be improved. In no-mobility scenarios, minor gains were achieved as per the research. Further, for full mobility and partial mobility conditions, important and substantial gains were achieved.

Janusz Henryk Klink, et.al (2017) studied few problems that were relevant to VoIP based dimensioning of selected networks [16]. Here, for the legacy circuit switched networks, the service quality measures were reviewed and the key performance indicators were applied. By including the average bandwidth that was needed to perform a one voice discussion, it was possible to calculate the capacity of complete network. Here, the research performed several analytical calculations. Finally, for the networks providing multimedia services, the directions of further research were described. Towards the video communication and consistent quality evaluation, particular attention has been given.

III. RESEARCH METHODOLOGY

The network traffic classification method is performed to categorize the traffic against malicious or non-malicious. This technique helps in predictive the malicious activities of active users. To categorize the network using proposed methodology, three important steps are applied. In the initial step, k-mean clustering method is applied that clusters the data against being similar and dissimilar. To refine the dataset given in the form of input, the redundancy and missing values are few problems that are eliminated. To calculate the central point of network, k-means clustering approach is applied in

the second step. The arithmetic mean of complete dataset is calculated here. The Euclidian distance is calculated from the central point such that the similar and dissimilar points are distinguished. Similar data points are included by one cluster. The data points in separate clusters are dissimilar. SVM classifier is applied in the last step such that the data points can be categorized among two separate classes. To improve the accuracy and performance of classification method, KNN classifier is used that clusters the un-clustered points. It also calculates the Euclidian distance and separates the similar and dissimilar kind of data.

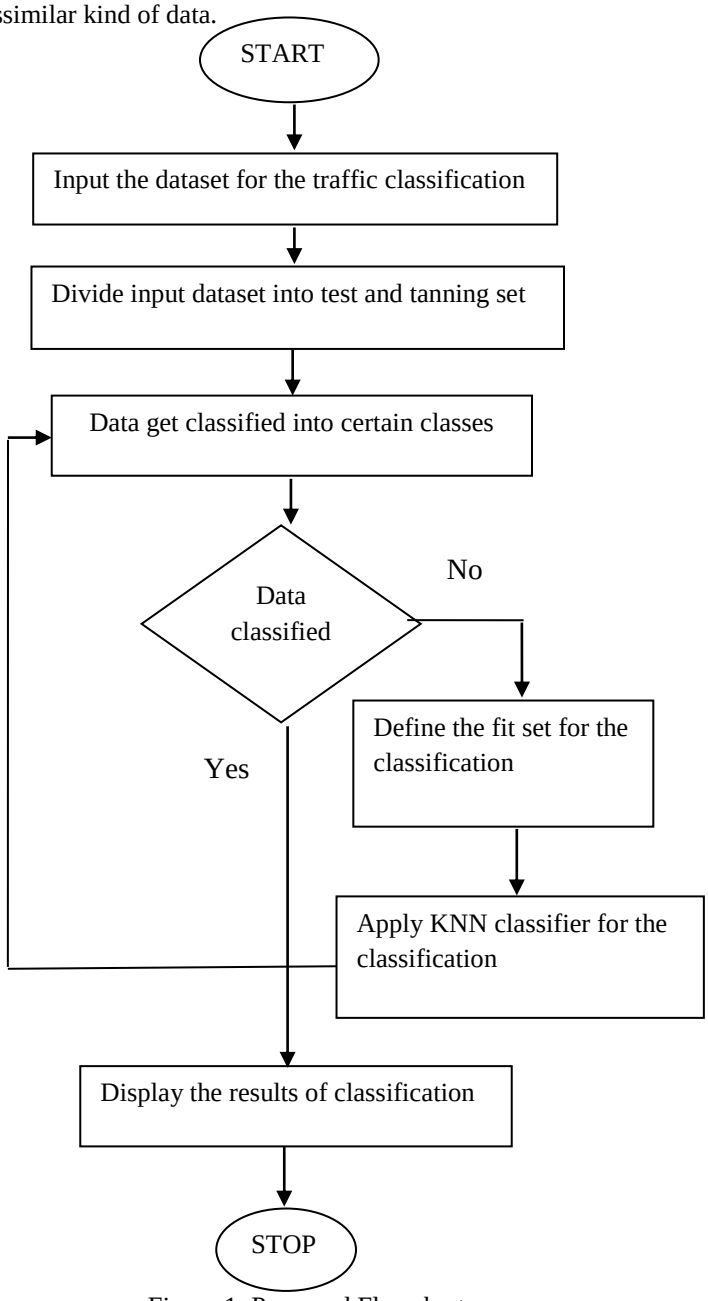

Figure 1: Proposed Flowchart

IV. EXPERIMENTAL RESULTS

The proposed research is implemented in Python and the results are evaluated by comparing proposed and existing techniques is terms of existing parameters.

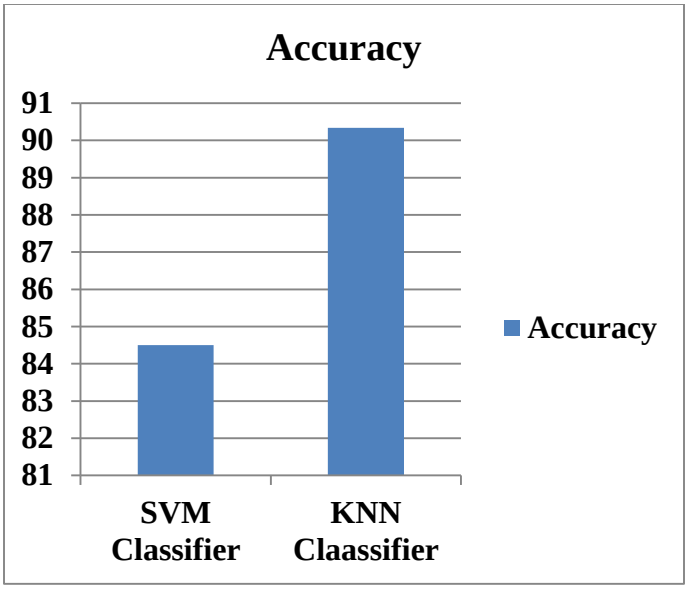

Fig 2: Accuracy Comparison

The performances of KNN and SVM classifiers when they are applied in network traffic classification are compared as shown in the above figure 2. The results show that higher level of accuracy is achieved by applying KNN classifier instead of SVM classifier.

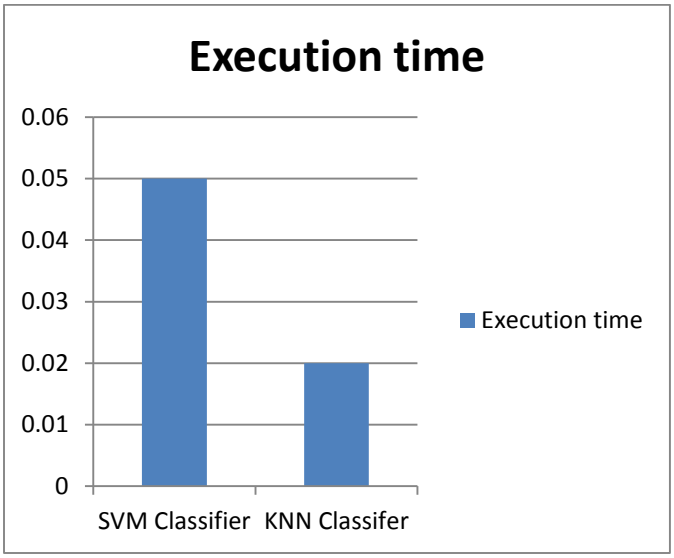

Fig 3: Execution Time

The performance results of proposed and existing algorithm are compared in terms of execution time as shown in figure 3. The comparison graphs show that the KNN classifier provides better outputs in comparison to SVM classifier.

V. CONCLUSION

Data classification is an important task to be performed in machine learning. This step is performed by the developed computer programs such that the labeled datasets can be achieved. Therefore, it also helps in predicting the unlabeled instances. Due to the availability of higher numbers of applications, various data classification systems are designed here. There are certain algorithms that are used very commonly in these fields. An important factor that is based on these algorithms is the appropriate parameter tuning. The numbers of neurons, hidden layers of ANN, value of k in KNN classifier are some of the calculations to be performed in this research.

VI. REFERENCES

[1] Seonghoon Moon, Juwan Yoo, and Songkuk Kim. Exploiting Adaptive Multi-interface Selection to Improve QoS and Cost-efficiency of Mobile Video Streaming. IEEE International Conference on Mobile Services, (pp. 134-141), 2015.

[2] Pablo Montoro, Eduardo Casilari. (2009). A Comparative Study of VoIP Standards with Asterisk. Fourth International Conference on Digital TelecommunicationsI, (pp. 1-6), 2009.

[3] J.Rosenberg, H.Schulzrinne, G.Camarillo, A.Johnston, J.Peterson, and R.Sparks, M.Handley, E.Schooler, "SIP: Session Initiation Protocol", Internet Engineering Task Force (IETF), Request for Comments (RFC) 3261, Jun.2002.

[4] Khaled Dassouki, Haidar Safa, Abbas Hijazi, and Wassim El-Hajj. "A SIP delayed based mechanism for detecting VOIP flooding attacks," IEEE International Wireless Communications and Mobile Computing Conference. Pp. 588-593, 2016.

[5] H. Sengar, D. Wijesekera, H. Wang and S. Jajodia "VoIP intrusion detection through interacting protocol state machines." In proc. of the Int. Conf. on Dependable Systems and Networks (DSN 2006), June 2006.

[6] Naktal Moaid Edan, Ali Al-Sherbaz, Scott Turner, and Surat Ajit. "Performance evaluation of QoS using SIP & IAXX VVoIP protocols with CODECS," SAI Computing Conference. Pp. 631-636, 2016.

[7] Andis Arins. "Latency factor in worlwide IP routed networks," IEEE 2nd Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE). Pp. 1-4, 2014.

[8] Yahya Z. Mohasseb, Mohamed T. Moubarak. "A QoS oriented analytical model for BGP multihomed networks," IEEE 16th Mediterranean Electrotechnical Conference. Pp. 31-34, 2012

[9] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, "QoE of VoIP Service: A Survey of Assessment Approaches and Open Issues," IEEE Comm. Surveys & Tutorials, vol. 14, No. 2, pp. 491-513, 2012.

[10] I. Martinez-Yelmo, I. Seoane, and C. Guerrero, "Fair QoE measurements related with networking technologies," WWIC 2010, LNCS 6074, Springer-Verlag Berlin Heidelberg, pp. 228–239, 2010.

[11] Mario A. Ramirez-Reyna, S. Lirio Castellanos-Lopez, Mario E. Rivero-Angeles, "Connection Admission Control Strategy for Wireless VoIP Networks Using Different Codecs and/or Codec Mode-sets", The 20th International Symposium on Wireless Personal Multimedia Communications (WPMC2017)

[12] Murizah Kassim, Ruhani Ab. Rahman, Mohamad Azrai A.Aziz, Azlina Idris, Mat Ikram Yusof, "Performance Analysis of VoIP over 3G and 4G LTE Network", IEEE, 2017

[13] Jan Holu, Michael Wallbaumy, Noah Smithy and Hakob Avetisyan, "Analysis of the Dependency of Call Duration on the Quality of VoIP Calls", IEEE, 2018

[14] Eko Ramadhan, Ahmad Firdausi,3Setiyo Budiyanto, "Design and Analysis QoS VoIP using Routing Border Gateway Protocol (BGP)", IEEE, 2017

[15] Mohammad Tariq Meeran, Paul Annus, Yannick Le Moullec, "Approaches for Improving VoIP QoS in WMNs", IEEE, 2017

[16] Janusz Henryk Klink, Tadeus Uh, "Quality-aware network dimensioning for the VoIP service", IEEE, 2017