IMPLEMENTATION OF MACHINE LEARNING FOR NETWORK TRAFFIC CLASSIFICATION

Ketan Gupta[1], Nasmin Jiwani[2], Md Haris Uddin Sharif[3], Vazeer Ali Mohammed[4], Murtuza Ali Mohammed[5], Mehmood Ali Mohammed[6]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

It may be necessary to detect which applications are moving via the networks inside the internet community in order to carry out specific tasks. Internet service providers (ISPs) generally employ network traffic classification to identify the prerequisites for a connection, It thus impacts the effectiveness of the cable network at the moment. Each one of the Internet Protocol (IP) methods—bandwidth, stream, and ML—has unique advantages and disadvantages. The Machine learning approach [5–9] is well-liked these days due to its vast use across disciplines and the growing knowledge among many researchers of its methodology when specifically compared to everyone else. Results from the Naive Bayes and K-nearest algorithms are then contrasted in this study when they are applied to a networking-specific dataset obtained utilizing live stream feeds and an Ethernet software. To develop a machine learning algorithm, the pandas and numpy arrays modules, the sklearn module for Python, and other help modules are used. Our research demonstrates that the K nearest method outperforms the Support Vector Machine, Nave Bayes, and Decision Tree algorithms in terms of efficiency.

Keywords: Decision trees (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Support vector machines (SVM).

[4,5]Research Scientist, Department of Computer and Information System, Lewis University, Illinois, USA
[1,2,3,6]Research Scientist, Department of IT University of the Cumberlands Williamsburg, KY, USA
Email: [1]ketan1722@gmail.com

DOI: 10.31838/ecb/2023.12.s3.078

1. Introduction

Network Activity Characterization places a lot of emphasis on problems related to new technologies. [1] discusses a number of identification techniques based on machine learning. examines the influence of the efficiency of a certain application protocol on the real quality of a broadband service provider. It has the ability to recognize any unknown networks that try to block a given traffic lane. In this way, they can discover more about its characteristics. One may also assess the dangers that a network may experience as a result of certain security breaches by using the aforementioned ability to recognize unknown networks. Regulating the network's architecture and quality of service (QoS) is also crucial, and these tasks may be accomplished if effective network categorization techniques are employed. We can prohibit or let particular network traffic if we classify our network effectively. In the end, networking classification raises the efficiency and revenue of the cable network. Over the past few decades, a number of network activity applications have been created, but they have all been able to classify network data. First, communications have been categorized using the Terminal Driver Distraction Classification approach, which employed terminals to classify every network. It was a really effective strategy at first. Port-based categorization approach is used in the results analysis in [2]. Ports were registered with the Internet Assign Number Authority in order to be categorized [3]. A quick explanation of Internet traffic categorization is also given in [4].

The study of their outputs is then studied using a variety of machine learning approaches [5 – 9]. The researchers at [11] have developed a revolutionary ML-based model for content grading, while [12] analyses Internet traffic. In order to classify Internet traffic, packets must be matched to the application from which they originated. Network management relies on traffic classification, which is used for things like traffic trying to shape, strategy routing, and packet filtering, among other things. Businesses use it for customer profiling, which gives them valuable marketing information, governmental organizations and scientists use it to research worldwide Internet trends. A single IP packet cannot easily categorize because the protocol headers do not include an application name. Due to P2P traffic, the communication port number was no longer a reliable way to distinguish between different traffic classes in the early 2000s. DPI (Deep Packet Inspection) is another widely used and accepted method for determining a packet's classification. There are privacy and computational costs to consider despite its accuracy. In addition, traffic encryption has rendered DPI obsolete [6]. Network management relies heavily on traffic classification.

In addition to network security, traffic visualization, and quality of service monitoring, it may be used for a number of other things as well. Over the past ten years, traffic categorization has changed quickly as a result of the emergence of peer-to-peer traffic. Researchers are continuously exploring for novel approaches to stay up with the Internet's dynamic nature. Second, payload-based algorithms were created, which examine packages from connected networks and identify protocols based on the research. This technique is known as "Deep Packet Filtering" since it analyses packets. Due to the high cost of system installation and the subpar performance it offers for encrypting communications, this solution has, unfortunately, failed [1]. Due to these flaws, machine learning is used, which has become more and more common in recent years due to its accuracy and effectiveness. Labeled classes are converted into models, which are subsequently checked for validity using precision. The paper's contributions are listed below. The optimal method for analyzing network traffic is then chosen after conducting a comparison evaluation of several algorithms using machine learning methods for a network information source. The characteristics are captured using a wire shark, trained using Python libraries, translated to a csv file format, then evaluated using those libraries to further assist in prediction and comparison analysis. Then, these statistics are contrasted [3]. Along with K-nearest Neighbor (KNN) and Naive Bayes (NB) techniques, we employ DT, NB, KNN, and SVM. We discover that the KNN technique performs better than the options exist for these applications.

2. Related Works

This study investigates internet traffic identification using ml algorithms.

Researchers are increasingly looking for IP traffic categorization methods that don't rely on "completely established" TCP or UDP control signals or package content interpretation. Utilizing traffic data to assist in the process of classification and identification is becoming increasingly popular. To categorize IP traffic, which combines IP networks and several data collecting techniques, Natural Language Processing (nlp) techniques are used [5]. 18 significant publications from 2004 to early 2007 are evaluated together with contextualization and motivation of ML methods used for Netflow segmentation. Based on the ML approaches they use and the important communities they are located in, these publications are classified and assessed. A set of fundamental requirements must also be met for ML-based traffic

classifications in functioning IP networks, and the evaluated works are ranked according to how effectively they meet these requirements. The organization also talks about current issues and barriers in the industry.

A comparison of low- and medium-intensity service assault dynamic queuing systems

Denial of Service (DoS) threats are posing a growing danger to the global inter-networking infrastructures. The crowded suggested controller for TCP is particularly resistant to a range of internet situations as a result of the underlying assumption of terminal collaboration. [12] For firewalls and neutralisation systems, low-rate denial-of-service attacks, particularly emerging threats, are more difficult to detect. In this paper, we examine these assaults. By utilizing analytical modelling, simulators, and network experiments, we propose that less DoS traffic situations that employ TCP's restoration length approach can decrease TCP streams to a fifth of the standard suitable price while avoiding detection. We study the intrinsic limits of randomized time-out techniques in preventing comparable low-rate Denial of Service (DoS) occurrences due to risks of protocol homogeneity.

Comparative Study of IP Traffic Classification Using Machine Learning Algorithms

Due to the dramatic increase in online bandwidth over the past few years brought on by the usage of a variety of online services, IP traffic categorization is becoming more and more relevant to broadband providers, as well as to other people and international organizations. Because random port numbers are used so frequently in incoming packets rather than well-known ports and because of a variety of cryptographic techniques, traditional Net flow categorization techniques like indirect packet filtering techniques for ports total count and bandwidth are no longer widely used [11]. Machine learning (ML)-based categorization is gaining popularity. For the purposes of this study, routine internet traffic data was collected using a packet capture tool and reduced using attribute selection techniques. These datasets were used in conjunction with RBF (machine learning), MLP (machine learning), Bayes Net to categorize IP traffic (machine learning) and C4.5 (machine learning). The findings of this study show that, with an accuracy of roughly 94%, Bayes Net & C4.5 are good ML techniques for categorizing IP traffic.

Using a "learning-based approach to document ranking"

The most pertinent documents are located using a query document to determine document similarity. Using a score function, document similarity methods have been shown to initially approximate semantic similarity between such a query and the documents. As according their similarity scores, documents are then ranked. There are three stages to the Text Tiling algorithm in the literature: tokenization into block of text units, scoring, and determining the subtopic boundaries. Throughout this article, we looked at two different methods for document ranking and compared the resul ts to those of a machine learning approach. In the first place, documents are ranked according to the tf-idf concept, which uses a standard score calculation. Second, the Text ling approach is used to rank documents. Strategies are an important Oriental publication text messages, that really have no paragraph breaks, using Text Tiles is already being implemented in a user interface for an information retrieval system. When summarizing documents, textiles are a considerably superior method than ordinary score computation. This enhances the system's retrieval performance. In this research, we compared and contrasted two methodologies. Our results also included the statistical machine learning techniques, which can be used to solve a wide range of information retrieval issues. Moreover (IR).

A brief summary of a few key articles on the topic of Internet traffic classification Informatics, both theoretical and practical

Traffic categorization is crucial for network administration. In addition to network security, traffic visualization, and quality of service monitoring, it may be used for a number of other things as well. Over the past ten years, traffic categorization has changed quickly as a result of the emergence of peer-to-peer traffic. Researchers are continuously exploring for novel approaches to stay up with the Internet's dynamic nature. 13 papers in all on the topic of traffic categorization and related topics were released between 2009 and 2012. Our findings demonstrate the breadth of modern traffic classification algorithms, as well as probable future routes for traffic classification research: the value of several levels of categorization, the need for experimental tests, and the importance of standardized traffic datasets.

3. Methodology

This study employs a variety of machine learning methods to anticipate traffic or categorize network data, such as email, browsing, and other types of traffic parentheses. The equation numbers should be consecutive within the contribution

3.1. *Naive Bayes Algorithm*

His strategy is based on the Naive Bayes method, which claims that practically all training data for learning are independent of one another because each class feature was individually evaluated when computing the whole. This theory outperforms many other Machine Learning approaches, which makes it particularly useful with a large training dataset. A thorough explanation of the Nave Bayes method may be found in [8]. Each of the n "k" attribute values is represented by the vector V, which equals the sum of the "k" values for each tuple in the training sample D. The "k" classes are C1, C2,..., Ck. According to Naive Bayesian classification, it only qualifies as a class Cy when its posterior distribution is bigger than class C1, C2,...Ck, where x y. Naive Bayesian classification states that an input pair I belongs to class Cx only when its posterior distribution is bigger than that of any other class Cy among C1, C2,...Ck, where x y, according to Naive Bayesian characterization

$$P\left[\frac{C_x}{I}\right] > P\left[\frac{C_y}{I}\right], \quad (1)$$

$$P\left[\frac{C_x}{I}\right] = \frac{P\left[\frac{I}{C_x}\right]P[C_x]}{P[I]}. \quad (2)$$

3.2. *K-Nearest neighbors*

This strategy is focused only on saving and finding new samples based on the specified parameter (mostly distance). The K-NN methodology has been used in statistical methods and predictive modelling for a very long period. Distances are calculated by the programme as a judgement parameter produces mediocre results when the feature data types are in fact quantitative & categorical. The test dataset approach may be used to assess machine learning-based algorithms. The results of the initial dataset are tested during the training phase. [4] One often used machine learning method is K-fold Serial Correlation. The steps in this procedure are as follows: In order to start this procedure, the first batch must be divided into k equal subgroups. To make things clear, we'll call these subsets as "folds," from f1 to fk being appointed to them. Repeat the loop for j = 1 to j = k. The remaining k-1 sets should be cross-validating training sets for confirmation, and the fold should represent the Authentication subset. By contrasting the validating findings with the actual values in the testing dataset, this cross-validation trained machine learning technique's correctness is assessed. The eventual efficacy of a machine learning designer is predicted by averaging the accuracy outcomes from k cross validation tests. The method's benefits and drawbacks are listed below. This strategy outperforms other ones when it comes to training with a large and noisy dataset. The optimum K value prediction is the most challenging aspect of this strategy. Because range is computed for each data entry, utilizing this approach has a significant processing cost. The inherent ambiguity that results from utilizing "length" as a criterion to evaluate effectiveness is another disadvantage.

3.3. *Decision tree algorithm*

By assuming the persistence criteria defined and decided throughout the phase of the training sample, a Decision Tree provides characterization that might predict category targeted parameter values by using the "training" concept. It has benefits since it is simple to explain [6] and it makes decisions much like a person would. However, the computations might get challenging if there are a lot of categorization IDs. Several techniques might be used to construct a decision tree. These include the ID3, CHAID, C4.5, and MARS, to name a few (Categorization of Regression Model). To create predictions, the CART technique is utilized. One well-known benefit of the Decision Tree technique is its transparency. It implies that the best method for attaining the best outcomes has been thoroughly examined before the findings are provided. It is a big advantage because of how unique the circumstance is and how much emphasis is placed on it. It can take into account any option due to its thoroughness. This tree is not only more aesthetically pleasing but also much more user-friendly. This works effectively even when working with records that contain both quantitative and qualitative information. These algorithms can become even more volatile with a slight modification in the input data, which could result in unsatisfactory results. When evaluated against a range of massive datasets, complex trees may perform worse. Because unbalanced datasets occur when algorithms detect perturbation in the information

3.4. *Support vector machine*

SVMs are a group of associated supervised learning techniques that may be used to solve a wide range of categorization and regression issues. Members of their linear categorization family are included. SVM's ability to reduce empirical categorization while concurrently improving geometrical margin makes it special. SVMs are hence sometimes referred to as Makes More Sense Classification models. SVM stands for structural risk minimization (SRM). The input vector may be used to build a hyper plane with the maximum separation using SVM [9]. Two parallel hyper planes are built along one side of the higher dimensional area seeking information. The distance between two perpendicular hyper planes gets wider as hyper - parameters are split. The

SVM is a productive technique for convex combinatorial optimization because it lacks local minima. Since SVM is based on the estimate of a test error rate cap, it appeals to the majority of analysts.

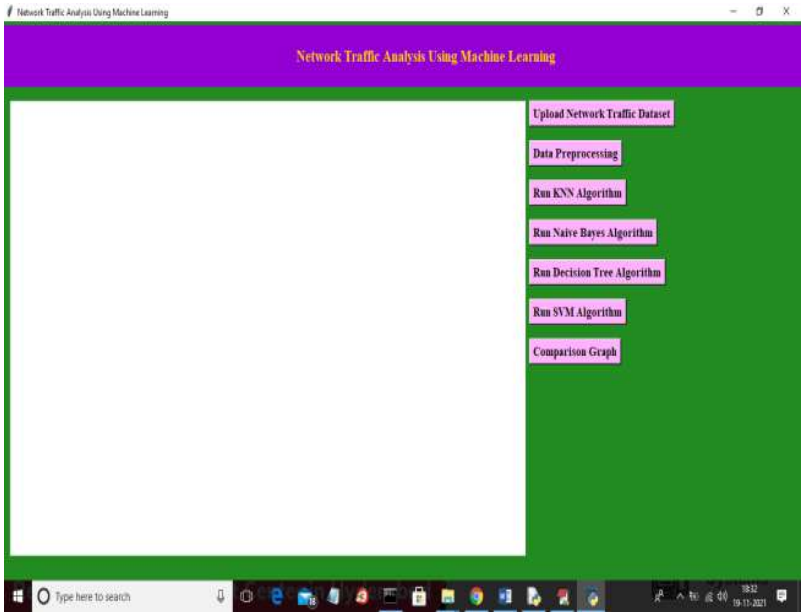
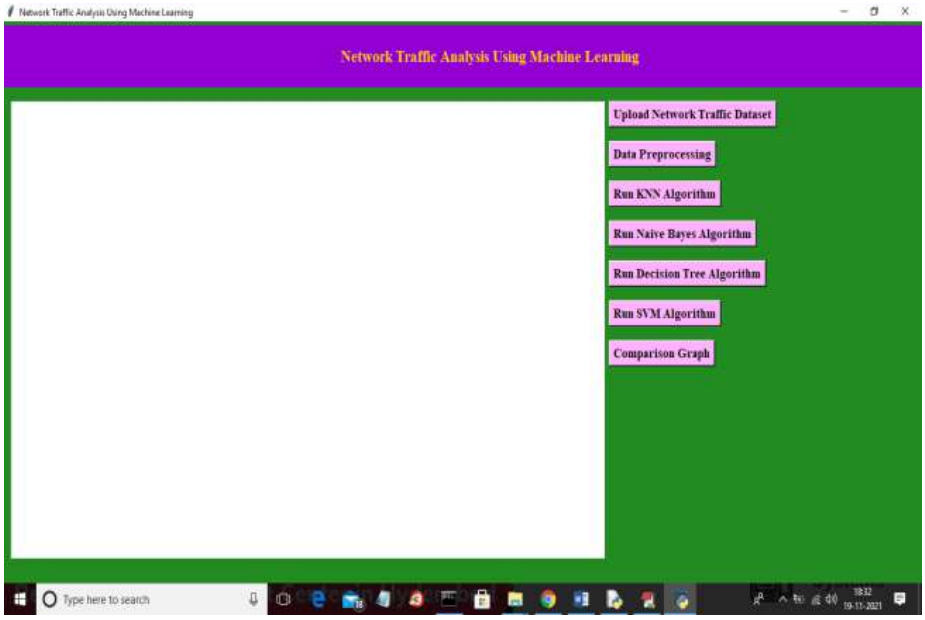


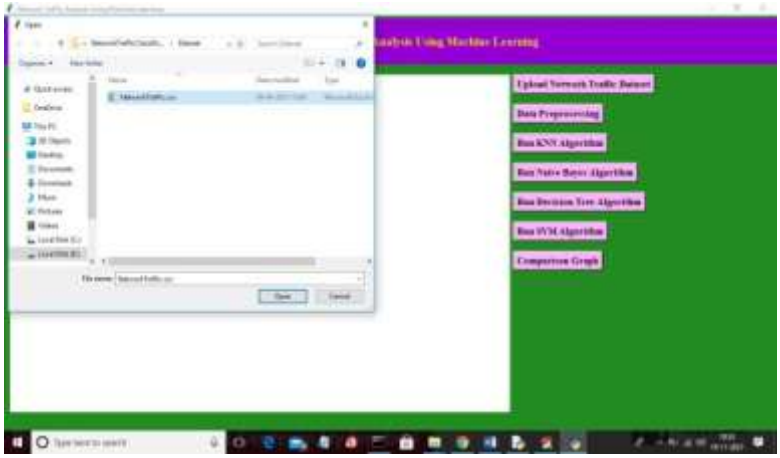
Fig. 1. Uploaded Network Traffic Database by clicking upload Network Traffic Dataset in the above result.

4. Result and discussion

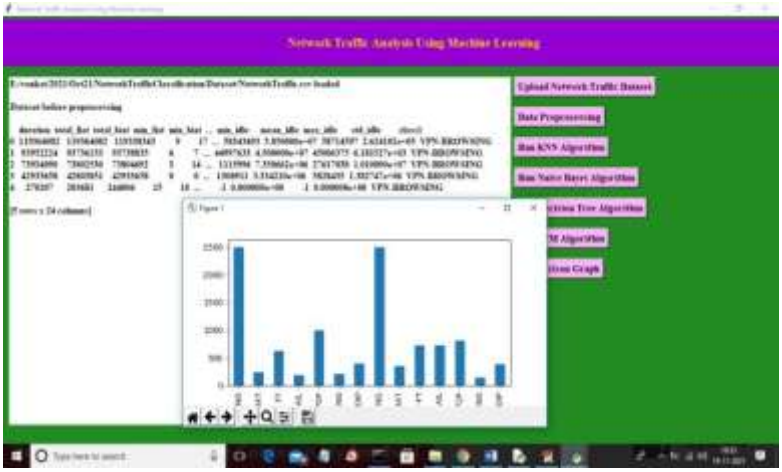
In this project we are using various ML techniques such as SVM, KNN, Decision Tree and Nave Bayes to predict traffic or classify type of network data such as BROWSING traffic, MAIL traffic etc. Lots of network traffic type of data is available but in this project we are training ML algorithms to predict or classify 14 different types of traffic. To run the project to get below result.



Uploaded Network Traffic Database by clicking upload Network Traffic Dataset in the above result



Choose NetworkTraffic.csv and then click "Open" to load the dataset. The outcome is shown below.

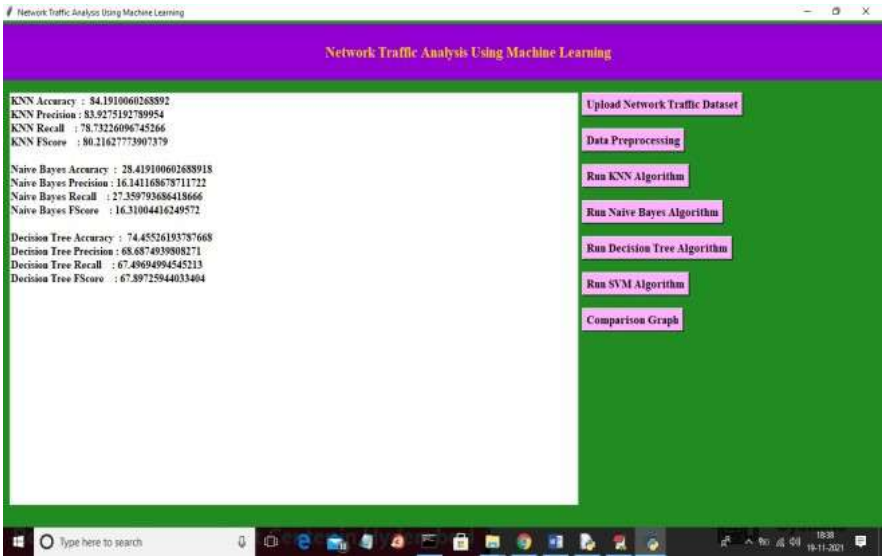


There are a lot of non-numeric values in the dataset that we need to analyse, therefore we can see that in the graph x-axis (traffic type) and y-axis (the number of entries in the dataset for that traffic) are shown. To tidy up the dataset, click 'Data Pre-processing' after closing the graph above.

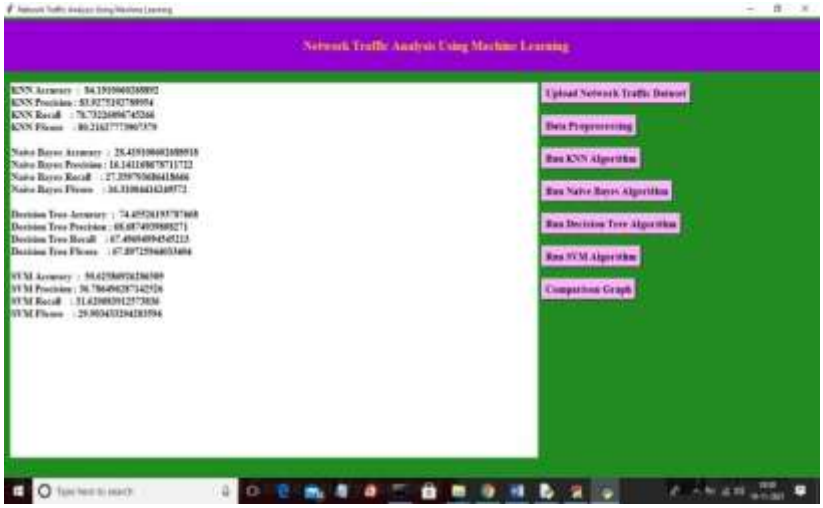
To summarise, we can see the total number of observations & fields found in the dataset, as well as a percentage breakdown of the dataset into train and test records, in the results shown above. Click the "Run KNN Algorithm" tab to start training KNN after all of the training and testing data has been gathered.

KNN accuracy is shown in the above results. With KNN, we achieved an accuracy rate of 84%, and the Run Nave Bayes Algorithm tab will be used to begin training the model.

Using naive bayes, we were able to acquire a 28 percent accuracy rate for the same dataset, but when we clicked the 'Run Decision Tree Algorithm' button, we were able to see the results below



In above screen for same dataset with SVM we got 74% accuracy and now click on 'Run SVM Algorithm' button to get below result



SVM Accuracy is shown in the results above. Using SVM, we were able to get a 50% accuracy rate. Click on 'Compare Graph' to see the following result.

This graph indicates that KNN outclasses all other algorithms on the x-axis and the y-axis. V. 3

5. Conclusion

This study investigates network traffic evaluation methods to better comprehend machine learning algorithms for dataset. The study may be quite helpful to analysts who are just starting out since it enables them to choose the ml model that is most appropriate for this specific method. The different ML approaches that will be learned in the next phase are first evaluated using the internet traffic separation. ML algorithms are used to categorize unknown actions and manage system performance. After that, the procedures are examined using machine learning techniques. Additionally, classifications based on this data are created using different machine learning algorithms, and their effectiveness is tested. We find that the KNN algorithm outperforms Support Vector methodology, Decision Tree and Nave Bayes Methodology along with high precision due to KNN's superior classifications to Decision Tree Method and Nave Bayes. among the other three techniques — DT, SVM and NB—we discovered that KNN was the most stable one while testing our training data set [5]. It is also feasible to keep the accuracy at its highest level.

6. References

Nguyen, Thuy TT, and Grenville Armitage. "A survey of techniques for internet traffic classification using machine learning." IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, 2008.

M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, "Network traffic classification techniques and comparative analysis using machine learning algorithms," in Proc. IEEE International Conference on Computer and Communications (ICCC-2016), pp. 2451-2455, 2016.

Internet Assigned Numbers Authority (lANA), http://www.iana.orglassignments/port-numbers, as of August 12, 2008.

Pawel Foremski, "On different ways to classify Internet traffic: a short review of selected publications Theoretical and Applied Informatics, " 2013.

K. Sing and S. Agrawal, "Comparative Analyssis of Five Machine Learning Algorithms for IP Traffic Classification," in Proc. IEEE International Conference on Emerging Trends in Network and Computer Communication (ETNCC-2011), pp. 33-38, 2011.

Q. Dai, C. Zhang and H. Wu, "Research of Decision Tree Classification Algorithm in Data Mining," International Journal of Database Theory and Application, vol. 9, no.5, pp. 1-8, 2016.

Cristina Petri, "Decision Trees", Cluj Napoca, 2010.

S. Karthika and N. Sairam, "A Naïve Bayesian Classifier for Educational Qualification," Indian Journal of Science and Technology, vol. 8, no. 16, Jul. 2015; DOI: 10.17485/ijst/2015/v8i16/62055.

D. K. Srivastava and L. Bhambhu, "Data Classification Using Support Vector Machine," Journal of Theoretical and Applied Information Technology, vol. 12, no. 1, Feb. 2010.

Wireshark tool: https://www.wireshark.org/docs/dfref/.

S. Patel and A. Sharma, "The low-rate denial of service attack based comparative study of active queue management scheme," in Proc. 2017 Tenth International Conference on Contemporary Computing (IC3- 2017), pp. 1-3, 10-12 Aug. 2017.

S. Patel, K. Khanna, and V. Sharma, "Documents ranking using learning approach," in Proc. 2016 International Conference on Computing, Communication and Automation (ICCCA-2016) , pp. 65-70, 29-30 Apr. 2016.