IJBPAS, June, 2018, 7(6): 999-1011

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

AN EFFECTIVE NEW APPROACH FOR NETWORK TRAFFIC CLASSIFICATION

RAFEEF FAUZI NAJIM AL-SHAMMARI

Department of Computer Science, College of Science, University of Kerbala, 56001, Karbala Iraq

Rafeef.fauzi@uokerbala.edu.iq

Received 4th Nov. 2017; Revised 6th March 2018; Accepted 24th March 2018; Available online 1st June 2018

DOI: https://doi.org/10.31032/IJBPAS/2018/7.6.4462

ABSTRACT

Now days the growth of Internet data has been increased, billions of the users are employed the Internet. Therefore, Managing and monitoring the internet become significant task of network administrators and developers. By Network traffic classification and identification, they can provide network management, security monitoring by block unwanted network traffic, network planning and Quality of Service (QoS). Internet traffic classification is the process of classifying network traffic according to various network applications into a number of traffic classes label. The network traffic classification can help to preserve smooth operation of the network traffic. Currently, the machine learning algorithms have widely implemented in network traffic classification for obtaining QoS of network. In this paper, I attempt to additional automate the process of classification network traffic. The statistics based network traffic is considered in my research work. In my presents research work, the proposed model is evaluated by using real-world network traffic traces that are also publicly available. The J48 algorithm is proposed to classify network traffic further enhance the network operations. In the evaluations, I as well compare my proposed model with existing methods. Experimental analysis demonstrates that the proposed model gives extremely good accuracy result as compare to other existing models.

Keywords: Network traffic classification, proposed model, Naïve Bays and Random Tree algorithms

1. INTRODUCTION

The Internet is getting to be focal in our life and work. In order to chatting in Face book to finding the cure for tumor, relatively every part of our life is some way or another identified with the Internet. Long gone are the days when the first ARPANET was conceived. From that point forward, the system has been in steady development changing the introductory ARPANET in what we today know as the Internet, a tremendous combination of interconnected PC systems. Regardless of its critical part in our life, the learning about its activity is a long way from being totally comprehended. The consistent presentation of new system designs, conventions and applications amid the most recent decades brought about a consistently developing substance hard to ponder and get it. This has impelled the exploration group to better break down the system activity and achieve some light the intricate task of the Internet.

In the previous couple of decades we have seen the astonishing development of the Internet as a world-wide communication infrastructure. We have watched its development from the soonest applications, for example, the primary email, up to huge circulated applications utilized by about two billion individuals on the planet [2]. On a fundamental level, the Internet is fabricated so the network system just gives an approach to data to achieve its last goal. The drawback of this effortlessness is that the network system neglects to address some of to days needs, particularly those identified with checking and administration of the network system. The need we address in this paper is the capacity to classify traffic streaming in the network system as indicated by the end-host application that has created it. This is, in fact, a challenging assignment, given that the network system does not straightforwardly offer such a usefulness, and that one needs to depend on restricted and regularly fragmented data to accomplish this objective. Especially, another field of study, as a rule alluded to as traffic classification, has wind up urgent for understanding the Internet. The classification of network traffic is a most significant area for studying, as well as has numerous vital applications for arrange administrators and IT managers. BitTorrent, Skype (i.e., P2P), Youtube, Netflix (i.e., gushing) or Megaupload (i.e., coordinate download) are a few cases of system applications that sooner or later totally changed the ideal models of the Internet. The classification of network traffic helps in a wide range of behavior. For example,

considering how new applications affect on the network system can help to better infrastructures, architectures or protocols. A precise classification can likewise enable Internet Service Providers (ISPs) to apply solid methods to apply Quality of Service (QoS) arrangements in light of the necessities of the applications (e.g., VoIP calls). At last, this opens another scope of charging conceivable outcomes for ISPs to take benefit of their foundations in light of their real uses.

Network traffic classification is the process toward partner network traffic streams with the network system applications that create them, which is additionally called application protocol identification. It fills in as an establishment for an extensive variety of exercises in systems administration, from network management system to network system security [1], from benefit separation to network traffic engineering, from slant investigation to network traffic research [3]. In this unique circumstance, the main object to classify is network traffic flow, which comprises of successions of network system packets traded between sets of endpoints with the end goal of between process correspondences over computer networks. The network classification can be founded on various data of the traffic flows, for example, port numbers, application payloads, statistical features of the network flows. In this paper, I talk about the difficulties looked by the present approaches for network traffic classification. I suggested new approach for enhancing network traffic classification. The J48 algorithm is more powerful for classification network applications.

2. Significant of Study

Network traffic classification understands the fine-grained perceivability of the sort of traffic crossing over the network systems. It is a standout amongst the most basic functionalities in the modern network. In order to management purpose, network system administrators are constantly excited about acquiring perspicuous mapping between each network system application and its relating network traffic, in order to get a handle on exact and extensive data about the applications utilizing their network systems. In view of the learning, they can implement administrative and security arrangements by forcing an arrangement of fine-grained rules in regards to the entrance to particular sorts of applications, services and substance. For instance, campus or enterprise network systems can oblige peer-to-peer (P2P) file sharing and gaming network traffic in order to save network assets for priority services or business basic

applications. Traffic classification is the center module of network system security including Firewalls and Network Intrusion Detection Systems (NIDS, for example, Snort [4] and Bro [5]. For instance, it empowers dynamic access control in versatile firewalls and application-level security observing and inspecting in NIDS, which identifies and anticipate unusual and abuse exercises because of vindictive clients, malwares, and Distributed Denial of Service (DDoS) assaults. As of late, ISPs are likewise being required by the legislatures to give capacities of Lawful Interception (LI) of Internet traffic, much the same as phone organizations need to help capture attempt of phone utilization.

3. Related Works

A numbers of research works has been done in the area of network traffic classification by application types. Several classification algorithms have been applied. In recent times machine learning algorithms are widely employed in this field. In the following section I will discuss some of the approaches done so far, which I have motivated this work:

Mohammad Reza et al. [6] used four variants of Neural Network estimator for classification network traffic applications. They have deceided to apply, Multilayer Perceptron (MLP); NARX (Levenberg Marquardt) and NARX (Naïve Bayes). It is observed that the accuracy of model was very good with compare with existing models. Li Jun et al. [7] presented TAN, C4.5, NBTree, Random Forest and distance weighted KNN algorithms for traffic classification scheme. In order to used genetic algorithm for feature selection for improving dimensionality reduction. From experimental results. It is observed that the C4.5 and Random Forest are superior algorithms for classification network traffic. Hamza Awad *et al*. [8] discussed various issue of machine learning algorithms for network traffic classification. The first issue characteristics of network traffic dataset. The second issue classes of real network data set. The third issue geographic place of network datasets. From experimental results obtained, it is observed that the classify online network traffic is better even the network data sets have more characteristics and captured from different geographic place.

Kalaiselvi1 *et al*. [9] introduced review literature of network traffic classification using machine learning. They study network traffic classification using different aspects such as using port based, payload-based and machine learning techniques. They found out that the machine learning techniques are

extremely significant for enhancing network traffic. With using machine learning for network traffic we can classify the traffic and enhance network security by blocking unwanted traffic. Ruchika Aggarwal *et al*. [10] Proposed Support Vector Machine (SVM) estimator for improving the network classification accuracy. Their methodology employed statistical feature based network traffic classification to improve feature discretization. From experimental results, demonstrated that the proposed has achieved much better classification performance than existing network traffic classification methods. Liu Zhen et al.[11] presented new Feature Selection method to improve internet traffic classification. They have suggested name of feature selection BFS. The BFS method is used to reduce features and alleviate multi-class imbalance in network traffic data sets. They have compared their new feature selection with correlation based filter and full feature network with using Naïve Bayes classifier. The results show that BF method performance better than all, the BF method used to preserve the balance of multi-class classification.

4. Methodology of Proposed Approach

This section describes the methodology used to evaluate the performance of my proposed model for network traffic classification. The big challenge of network traffic classification are encrypted applications of traffic and user data privacy in network. Machine learning algorithm is used to solve this problem for improving the performance of network traffic. First, the tool used for the evaluation is presented and then, the dataset used as ground-truth for the evaluation is described. I have suggested to use J48 algorithm for increasing the accuracy of network traffic applications. Various existing algorithms are implemented using Weka too for comparison with the proposed model. It is observed that the proposed outperformed of all the existing model. The detailed description of methodology of the proposed research is presented in the subsequent subsection.

4.1 Data set

Throughout my research work, I have collected data the high-performance network monitor

described in [12].With its loss-limited, full-payload data was captured with time stamp. They have captured data with different tome stamps period 24-hour period. The data was collected from institutions on site. There are about 1000 researchers, administrators and technical staff (users) in that institutions. The institutions are

connected by a full-duplex Gigabit Ethernet link. The traffic was monitored for each traffic set for bidirectional traffic. Table 1 shows classification example of data with different applications in data set.

Table 1: Classification example of data with different applications in data set

Classification Examples	Applications
BULK	ftp-control, ftp-pasv, ftp-data
DATABASE	postgres, sqlnet oracle, ingres
INTERACTIVE	ssh klogin, rlogin, telnet
MAIL	imap, pop2/3, smtp
SERVICES	X11, dns ident, ldap, ntp
WWW	www
P2P	KaZaA, BitTorrent, GnuTella
ATTACK	Internet worm and virus attacks
GAMES	Microsoft Direct Play
MULTIMEDIA	Windows Media Player, Real

4.2 J48 algorithm

Classification is supervise learning algorithms, it uses to build a model of classes from data set that contain class labels. Decision Tree is one of the classification algorithms. Decision Tree algorithm is process to discover the way the features-vector behaves for a number of instances in the data sets. The Decision Tree algorithm is used to generates the rules in order to predict the target variable. Decision Tree algorithm apply to classify grave distribution of the data set is easily comprehensible. J48 algorithm is one types of Decision Tree algorithms. J48 algorithm is an extension of ID3 algorithm. The further features of J48 are used to classify data sets have different issue for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc.

A. Basic Steps in the Algorithm:

1- The tree represents a leaf when the instance of attributes belongs to same class label. Then, the leaf is represented by labeling with same class label.

2- The algorithm is used entropy to calculate the information of attributes. When the attributes have tested, then the information gained method used for calculating each attributes to find out the best attribute. The best attribute is used as root of tree.

3. **Counting Gain**

This information gain is process employ entropy concept. The information gain uses to select the attribute have highest information for classification purpose. The Entropy $(\bar{t})$ of is calculated by:

$$Entropy\ (\bar{t}) = -\sum_{j=1}^{n} \frac{|t_j|}{|\bar{t}|} \log\ (\frac{|t_j|}{\bar{t}}) \quad (1)$$

$$Entropy\ (j|\bar{t}) = \frac{|t_j|}{|\bar{t}|} \log\ \frac{|t_j|}{|\bar{t}|}) \quad (2)$$

$$Gain\ (\bar{t}, f) = Entropy(\bar{t} - Entropy\ (j|\bar{t}) \quad (3)$$

4. Pruning tree

Tree pruning is most significant steps in the algorithm. It uses to remove outliers in the training data. Due to at time of classification there are some instance in all data sets which are not belong to any class. These instances are not well define and also different from other instance in dataset. The pruning tree is applied to decrease classification errors. Pruning is used to generate the tree.

5. Features of the Algorithm

(i) The J48 algorithm can handle both the discrete and continuous features. The algorithm uses threshold value for deciding types of features in the dataset.

(ii) The J48 algorithm can handle the missing values in the training data.

(iii) The J48 algorithm is used pruning of the tree for removing branches that are not assisting in reaching the leaf nodes of tree after constructed tree.

6.**Experimental Results and Analysis**

In this section my main object to evaluate the performance of my proposed model. To carry out my the experiment I have used 64 windows 10 Ultimate with the core i5 processor and 4 GB RAM. The proposed model for network traffic classification is implement by using Python programming. The 10-fold cross validation has applied for testing the performance of the proposed system. Based on experiment, I have implemented the following algorithm such as Naïve Bays and Random Tree algorithms for further comparison with the proposed model. These algorithms are implemented by using Weka tool data mining tool. The Accuracy, Recall, Precisions performance measure are employed to evaluate the proposed model. The original data set of network traffic classification contains 249 attributes with 24863 instances. It contains 12 classes of different network applications. Table 2 contains information about the network applications of the network data.

At the beginning of the analysis network traffic, the J48 algorithm is applied with using python programming. Table 3 shows the accuracy results obtained from the

proposed model and existing models. The J48 algorithm performed on real network traffic. All the applications are classified with effected instance values. The FTP-Data and Database applications give maximum 100% accuracy results has more effective classified as compared with existing Naïve Bays and Random Tree algorithms in terms of accuracy metric. It is observed that the proposed model give highest accuracy in all the network applications. Figure 1 displays the performance of the proposed model as compared with existing models for network traffic classification. It is investigated that the performance of the model better than all the another algorithms. Similarly, the proposed model give best performance with respect to Recall and Precision metrics. Table 4 and table 5 show the Recall and Precision results of my proposed model as compared with conventional algorithms. However, the Mail, FTP-Control, Database and FTP-Data network applications provide the best performance 100 % accuracy results in the terms of Recall metrics. Furthermore, Mail, Database, FTP-Control and Interaction traffic applications are classified extremely accurately and the proposed model obtained 100% precision results which are the assuring results. Figures 2 and 3 illustrate the Recall and Precision results of proposed model as compared with Naive Bays and Random Tree existing algorithms. It is observed that the proposed model is classified very effectively overall the conventional algorithms. The proposed model has outperformed all the other models.

Table 2: Information of Applications in data set

Network applications	Total instances
WWW	18211
Mail	4146
FTP-Control	149
FTP-PASV	43
P2P	339
Database	238
FTP-Data	1319
Multimedia	87
Services	206
Interaction	3
Games	0
Attack	122

Table 3: Accuracy results of proposed model with compared existing models.

Applications	The proposed model (J48) algorithm	Naïve Bays algorithm	Random Tree algorithm
WWW	99.93	93.27	99.48
Mail	99.97	97.56	99.27
FTP-Control	100	97.98	99.63
FTP-PASV	83.72	97.67	65.11
P2P	99.41	76.99	99.41
Database	100	98.31	97.47
FTP-Data	100	98.63	99.62
Multimedia	94.25	97.70	75.86
Services	99.02	99.02	88.46
Interaction	66.66	66.66	00.0
Games	0.00	0.00	00.0
Attack	90.37	80.32	58.19

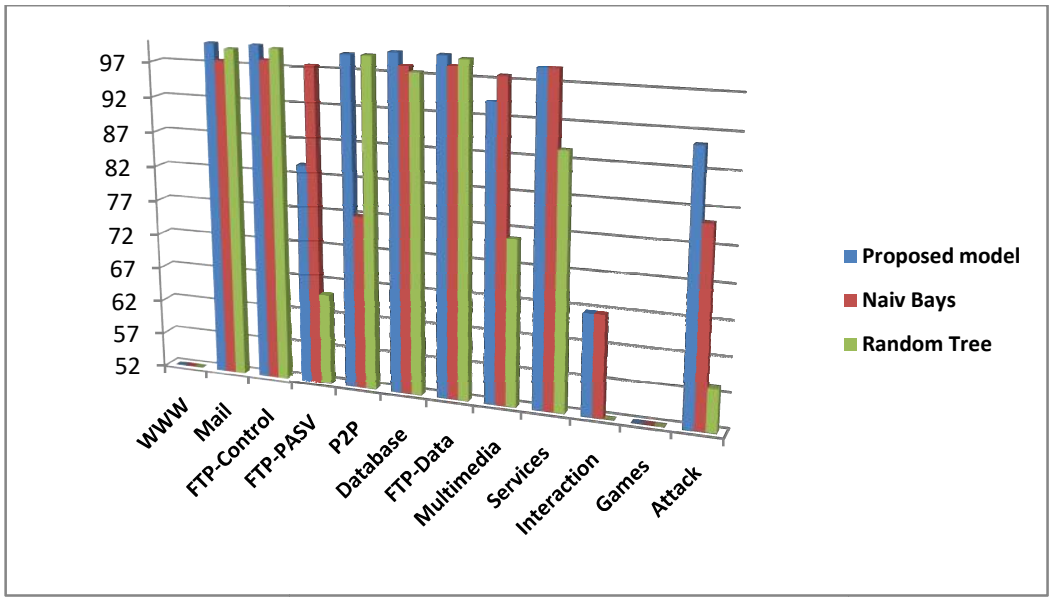

Figure 1: Performance of accuracy results

Table 4: Recall results of proposed model with compared existing models

Applications	The proposed model (J48) algorithm	Naïve Bays	Random Tree algorithm
WWW	99.09	93.03	99.05
Mail	100	97.07	99
FTP-Control	100	98	94.06
FTP-PASV	83.07	97.07	65
P2P	94.09	77	86.07
Database	100	98.02	97.05
FTP-Data	100	98.06	99
Multimedia	94.03	97.07	75.09
Services	99.90	99.09	99
Interaction	66.07	66	0.000
Games	0.00	0.000	0.000
Attack	90.04	80.03	58.02

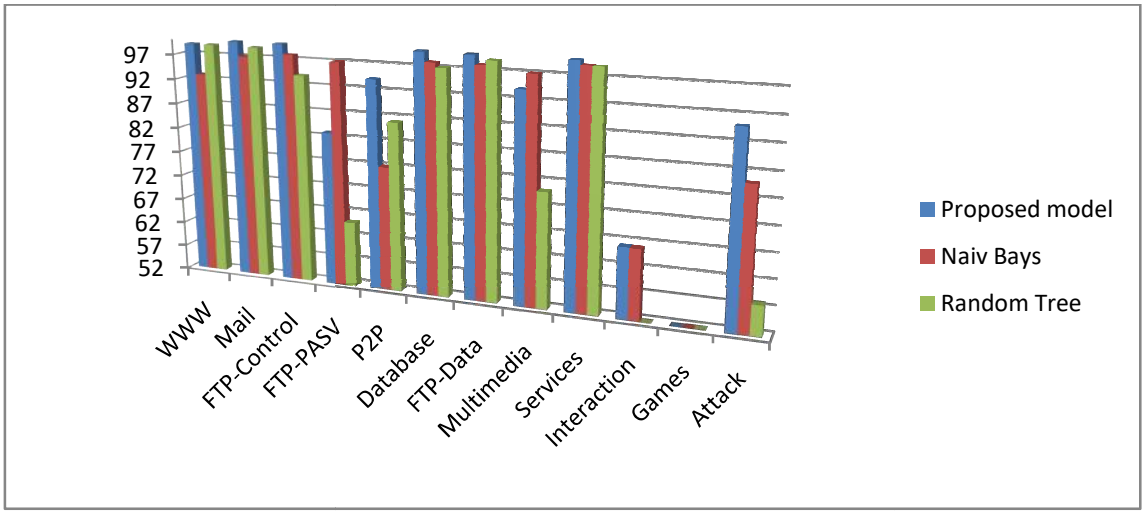

Figure 2: Performance of Recall results

Table 5: Precision results of proposed model with compared existing models.

Applications	The proposed model (J48) algorithm	Naïve Bays algorithm	Random Tree algorithm
WWW	99.07	98	99
Mail	100	99.07	99.03
FTP-Control	100	64.09	92.02
FTP-PASV	83.07	76	75.07
P2P	97.07	80	89.09
Database	100	98.03	93.03
FTP-Data	99.09	99	99.05
Multimedia	95.03	40.05	82.04
Services	99	94.09	98.06
Interaction	100	24	0.000
Games	0.000	0.000	0.000
Attack	93.02	10.03	55.09

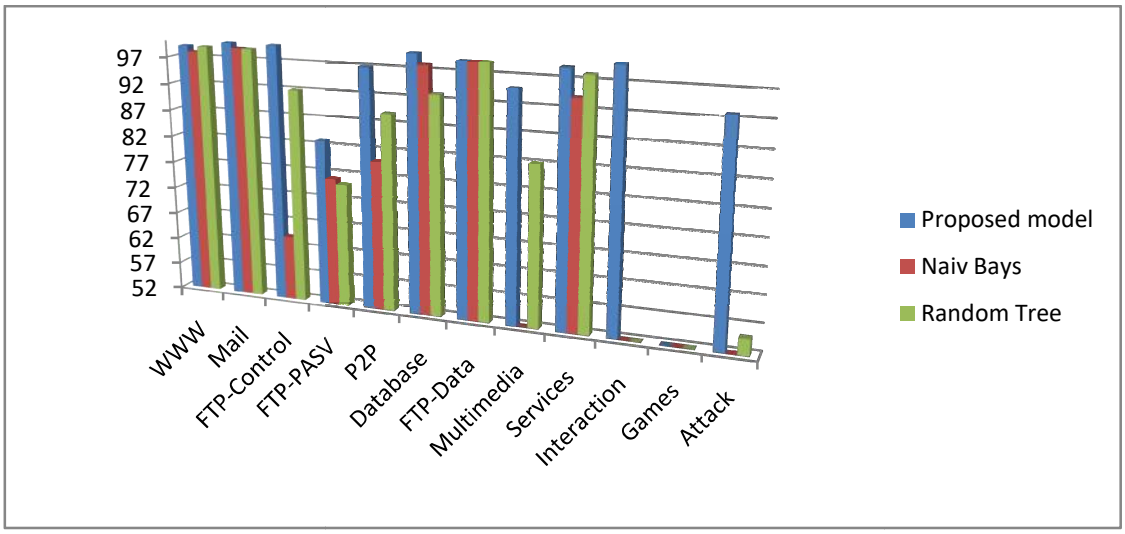

Figure 3: Performance of precision results

7. CONCLUSION

The capability to accurately make out the network traffic is a extremely significant issue to attain accurate and fast management tasks. Machine learning algorithms plays vital role in network traffic classification for enhancing QoS. In this paper I have demonstrated the powerful of the proposed model for classification network traffic. The performance of my proposed model is very promising for network traffic classification. The experiment results showed that the proposed model is able to enhance the network traffic classification. This research work I have applied J48 classifier without any preprocessing approaches for improving the algorithm. The J48 algorithm is more powerful classifier for network traffic classification. Furthermore, all the applied machine learning algorithms obtain extremely effective performance results, however, I observed that the J48 algorithm have extremely effective performance as compared to existing Random Tree and Naïve Bays algorithms. In my experiment the proposed model obtain extremely efficient performance with respect to Accuracy, Recall and Precision classification measurements. In future, I will try to use different types of datasets.

REFERENCE

[1] Alberto Dainotti and Antonio Pescapé, "Issues and future directions in traffic classification", IEEE Network, vol.26, (2012), pp.35-40.

[2] P. Cheeseman and J. Strutz. Bayesian Classification (AutoClass): Theory and Results. In Advances in Knowledge Discovery and Data Mining, AAI/MIT Press, USA, 1996.

[3] J. Padhye and S. Floyd. Identifying the TCP Behavior of Web Servers. In Proceedings of SIGCOMM 2001, San Diego, CA, June 2001.

[4] N. Cascarano et al., "Comparing p2ptv traffic classifiers," in Proc. IEEE ICC, 2010, pp. 1–6.

[5] C. Bishop et al., Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 3, pp. 226–239,Mar. 1998.

[7] Mohammad Reza Parsaei, Mohammad Javad Sobouti, Seyed Raouf khayami, Reza Javidan.

Network Traffic Classification using Machine Learning Techniques over Software Defined Networks. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.7, 2017, pp .22-225

[8] Li Jun, Zhang Shunyil, Lu Yanqing and Zhang Zailong. Internet Traffic Classification Using Machine Learning. Communications and Networking in China, 2007. CHINACOM '07. Second International Conference on IEEE. 07 March 2008, pp.1—5

[9] Hamza Awad Hamza Ibrahim, Omer Radhi Aqeel AL Zuobi, Marwan A.Al-Namari, Gaafer Mohamed Ali, Ali Ahmed Alfaki Abdalla. Internet Traffic Classification using Machine Learning Approach: Datasets Validation Issues. Conference of Basic Sciences and Engineering Studies (SGCAC) , 2016 IEEE,PP-158—166

[10] T.Kalaiselvi, P.Shanmugaraja. Internet Traffic Classification Using supervised Learning Algorithms–A Survey. International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 04 | Apr-2016, PP.91-93

[11] Ruchika Aggarwal, Nanhay Singh. A new Hybrid Approach for Network Traffic Classification Using SVM and naïve Bays Algorithms. International Journal of Computer Science and Mobile Computing. Vol. 6, Issue. 6, June 2017, pg.168 – 174

[12] Liu Zhena, Liu Qiong. A New Feature Selection Method for Internet Traffic Classification Using ML. International Conference on Medical Physics and Biomedical Engineering. Elsevier 2012, 33 (2012) 1338 – 1345

[13] Wei Li, Marco Canini, Andrew W Moore, and Raffaele Bolla. Efficient application
identification and the temporal and spatial stability of classification schema. *Computer Networks*, 53(6):790–809, 2009.

[14] Zhang, Jun, et al. "Network traffic classification using correlation information." IEEE Transactions on Parallel and Distributed Systems 24.1 (2013): 104-117.

[15] S. Zander, H. Nguyen, G. Armitage, "Automated Traffic Classification

and Application Identification Using Machine Learning," Proceedings of The IEEE Conference on Local Computer Networks 30th Anniversary, Washington: IEEE Computer Society, 2005, pp. 250-257.

[16] J. Erman. M. Arlitt, M. Anirban. Internet Traffic Identification using Machine Learning Techniques: Proc. Of 49th IEEE Global.