Grenze International Journal of Engineering and Technology, July Issue

Online Hybrid Classifier System of Internet Traffic based on Machine Learning Approach and Port Number

Hamza Awad Hamza Ibrahim[1], Omer Radhi AL Zuobi[2], Awad M .Abaker[3] and Marwan A.Al-Namari[4]
[1-4]College of Computer at Al-Gunfudah, Umm Al-Qura University, Al-Gunfudah, Saudi Arabia
haibrahim@uqu.edu.sa, orzubi@uqu.edu.sa, amabker@uqu.edu.sa, masnmas2@uqu.edu.sa

***Abstract*—Internet traffic classification is valuable mechanism in the direction of traffic detection and monitoring. Even though several classification approaches were proposed by the research community, there still exist many open problems on Internet traffic classification. The hybrid classifier is a classifier which combines more than one classification method to identify the Internet traffic. Using only one method to classify Internet traffic poses many risks. Therefore, this paper proposed a hybrid classifier (HC) system to identify internet traffic. HC is based on two common classification methods, i.e. port-base and ML-base. CH was able to perform an online classification since it able to identify the live Internet traffic at the same time as when the traffic was generated. HC was used to classify three common Internet applications classes i.e. web, WhatsApp, and Twitter. HC is produces more than 90% classification accuracy which is higher when compared with other individual classifiers.**

***Index Terms*— Internet Traffic Classification; Machine Learning; Classificatiom Methods; Port-based method; Hybrid classifier.**

I. INTRODUCTION

Internet Server Provider (ISP) and network operators are usually interested in knowing the traffic carried in their networks for the purpose of optimizing network performance and security issues. Therefore, network traffic classification is an important foundation of identifying unknown Internet applications which have abnormal behaviors. In particular, network classification can detect traffic which includes threats such as denial of service, flooding attack and other such threats.

With the increase in Internet usage, a lot of Internet applications have been developed. However, these new applications can carry abnormal Internet traffic, which has a negative effect on the network performance. Some of the Internet applications generate several types of versions with different traffic attributes. For instance, online games usually constitute a huge number of the overall games over the world each year. Therefore, network traffic classification is very valuable to identify these large applications. Network traffic management is another issue which shows the importance of Internet traffic classification. When network managers plan to control network users through fair usage of bandwidth, they need first to know which type of applications they are dealing with. Thus, the managers cannot achieve their administrative tasks, unless they classify the network traffic. In the home network, traffic classification can help to enhance Quality of

Grenze ID: 01.GIJET.6.2.6_3

Service (QoS) of Internet services. In general, the identification of Internet traffic helps in different network management activities, such as bandwidth control, traffic engineering, fault diagnosis, application performance and anomaly detection [1].

A. Classification Methods

Port-based classification method is based on 16-bit port numbers on transport layer, which consist the information of source and destination ports. Simply, the classifier uses these port numbers to determine the application classes. In other words, the classifier read the port number from Internet Assigned Numbers Authority (IANA) and then uses this number to distinguish between the Internet applications' types. Port-based classification method has the following advantages: (i) it is very simple, (ii) it can be used to limit the worm's traffic, (iii) it is very fast, (iv) it can be applied by all routers and layer 3 switches, and (v) it is efficient in classifying protocols carried by fixed port number [2]. However, this method was not sufficient to classify the new internet applications during using of unknown-port number [3] [4] [5].

Payload-based classification or Deep Packet Inspection (DPI) is an individual packet inspection looking for unique signatures. This means that the packets will be investigated one by one to find a unique signature. This helps in knowing the packet that belongs to a particular class (application). According to researchers [6-8], payload based classification methods achieve higher accuracy. This is generically due to the fact that the unique signature (if it exists) always tells the truth with nothing to hide. The type of signature used in the classification of traffic is based on Internet application type and can be found in application or transport layers. The classifier can use the signature in text (string) or hexadecimal (HEX) formats. The classifier uses these signatures to decide which packet/flow belongs to which application.

Another method is machine learning (ML)-based or flow statistical-based and it uses a collection of information to classify the network traffic. The main advantage of this approach is that it can be used at any point in the network [2]. Unlike port and payload based methods which are based on specific port number and unique signature, statistical-based method can identify the traffic based only on statistical features calculated from network flows. Machine learning (ML) is the most common technique used for statistical based classification. Machine learning is one of the modern application classification techniques, which uses Artificial Intelligence to identify IP traffic. ML provides better solution in extracting real information from application features [9]. Moreover, some of ML algorithms are suitable for Internet traffic flow classification at high speed [10]. ML technique is performed in several steps; firstly, selection of a dataset which contains all or some of the features values. These features are attributes of traffic flow, such as packet length, inter arrival time, protocol, idle time and other such attributes. Secondly, application of the training stage for ML to establish classification rules; this is based on statistical computation extracted from the features. Lastly, application of the ML classification to unknown packets using the training rules from the second step. Due to the rapid nature of real time applications, important issue that must be considered when classifying Internet applications is the time of collecting the statistical values (to build the rules), which is assumed to be very short. ML consists of different algorithms categorized into two main types supervised learning and unsupervised learning.

Another classification method is to use hybrid method. Network hybrid classifier is defined as a classifier which uses more than one classification method. Port-based, payload-based, statistical-based and hardware-based are the common methods that are used in building hybrid classifiers. Each of the classification methods has some advantages as well as some limitations. The hybrid multistage classifier makes full use of the advantages of each of the partial methods [7]. However, the disadvantage of hybrid classifier is the complexity involved when using more than one stage. This complexity can be evaluated through the classification time versus classification accuracy. In other words, what is the trade-off between complexity and accuracy (which is expected when more than one method is used) than can be achieved

This paper proposed Hybrid Classifier (HC) based on two of pervious method, machine learning and port number.

B. Online Classification

In online classification, the decision about which packet (or flow) belongs to which particular class is based on the traffic speed. This is the same as any hardware classifier (Packet Shaper, SANGFOR) which is installed on the network path in order to classify the traffic at the network speed. Therefore, the online classifier is normally installed inline with the switch/router to identify the total traffic that passes through this device. Online classifier is very important to manage threats on the traffic such as denial of service, flooding attack and other similar threats. Most of the current classification methods do not support online

classification [11]. Furthermore, most of the published articles only focused on the classifier accuracy with the classifier trained based on the full flow. However, this approach cannot be implemented successfully for online classification [12]. One of the problems in online classification is the high traffic speed. The challenge will be to do all the following steps in the case of high network traffic speed, i.e. i) capturing the traffic ii) dividing it into flows iii) calculating the statistical features or checking the payload.
Offline classification is not helpful for online management and control mainly due to the performance reason [13]. Online network traffic classification is very important because of several reasons such as:

- Online classification is the basis to manage the real time network traffic. Therefore, in order to manage and control the Internet traffic, there is a real need for online classification.
- Online classification helps to prevent network threads and abnormal behaviors such as denial of service, flooding attack and other such threats.
- Developing of effective software-based online classification algorithms helps to reduce the use of hardware classifier (such as Packet-shaper) which has very high cost.

One of the goal of this paper to differentiate between "real Internet traffic" and "online Internet traffic". There is a big difference between these two terms; real traffic can be defined as any real Internet traffic which is captured in any network level, and at the current time this is not live traffic. While online traffic means the traffic which is currently running in the network (live traffic). Figure 1 illustrates the difference between real traffic and live traffic. Real traffic is more extended definition than live traffic. In other words, the live online traffic is a real traffic, but the real traffic is not always an online traffic. In the same manner, there are a big difference between online classification and real traffic classification. Real traffic classification is the identification of the real network traffic which can be called offline classification. This paper defines the online classification as a system which can receive and classify the Internet traffic at the traffic running time.

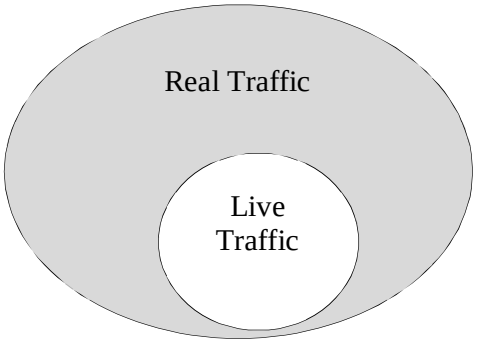

Figure 1 Real traffic and live traffic

II. Related Works

This section discusses some of the previous Internet traffic classification work. There are several research papers that have considered ML classifier which was used to classify a datasets in different ways such as packet traffic features, flow traffic features, statistical packet features, etc. [14], made a comparison between a five ML algorithms (MLP, RBF, C 4.5, Bayes Net and Naïve Bayes). The authors developed a real Internet traffic dataset which includes seven applications: web, e-mail, web media, P2P, FTP data, instant messaging and VoIP. In their work, they used Wireshark as the capturing tool. The result shows that, in the case of full features dataset, Bayes Net classifier provides the better accuracy of 85.33 %; when the authors applied the approach of reduced features, C4.5 provided the higher accuracy of 93.66%.
In [15], the authors proposed re-sampling methods to alleviate the data skew for network traffic classification. For the purpose of comparison, the authors used three types of sample datasets: stratified sampling, uniform sampling, and tuning sampling. The dataset includes nine classes which are: www, mail, bulk, attack, chat, p2p, multimedia, VOIP, and interactive games. Each class includes some of Internet applications (such as, www includes Web browsers, web applications, and IMAP). The applied methods is tuning sampling in order to maintain the accuracy, in other words, re-sampling of training data to decrease the data skew. What make this study important is the high number of applications which considered. However, the authors mentioned that the collected datasets includes traffic of thousand local users. This

indicates that the training and testing date sets was collected from different network levels. Thus, how the authors ensured that all the switches/access-points traffic has the same characteristics.
In [16], based on the analysis of P2P traffic classification technologies, a combining a packet-level classifier and a flow level classifier is proposed. The first level is a deep packet inspection based classifier at which work at the packet level to identifies the specific P2P traffic. The second step is a machine learning approach which classifies the remaining unknown P2P traffic at the flow level.
Bujlow et al. (2012) proposed a classification method based on the C5.0 ML algorithm. The authors recruited volunteers from the users to generate the real labeled traffic. Some software was installed on the volunteers' computers to capture the relevant traffic and submit the datasets to the server. C5.0 ML algorithm was used as statistical classifier to distinguish between seven types of applications (Skype, FTP, torrent, web browser traffic, web radio, interactive gaming and SSH). It is greatly acceptable that the authors developed a classifier to be network-dependent, which means it will train in each network independently. However, the traffic flows were collected from volunteers' NIC; the characteristics of this traffic can change when passing through network switches. In addition, the online classifier normally installed in the switch/router to identify the total traffic passes through this device.
The authors in [17] proposed an algorithm (named Skype-Hunter) to identify Skype traffic. The proposed algorithm is based on both signature-based and statistical traffic features. The experimental part of this works considered different scenarios like: 1) no restrictions on the transport protocols; which mean to use the direct connection between the Skype clients. 2) Presence of a NAT IP; this means the use of the IP network address translator which is a router function that can be configured to allow the addresses of a stub-domain to be reused by any other stub-domain. 3) Presence of a firewall which does not allow the use of UDP.
The authors of [18] a hybrid approach to classify network traffic using SVM and NAÏVE Bayes algorithms. The paper uses flow statistical feature to enhance feature discretization..

III. METHODOLOGY OF THE PROPOSED HC

This paper proposed hybrid online Internet traffic classification model based on machine learning technique and port number. Figure 2 explains in a simple way the ML classification system. The training stage is the main input and the classification result is the output of this system. If the input is valid, this means the output should be valid. On the other hand, the using of port number in classification was still helpful and it can be relevant for certain type of internet application traffic [19].

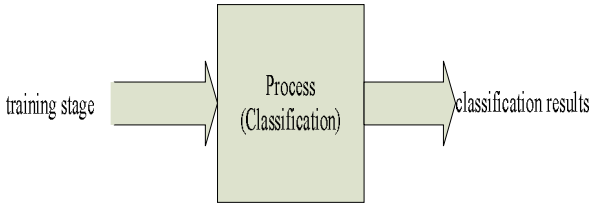

Figure 2 Classification system stages

Figure 3 illustrates the network environment of the proposed system which shows that the access point (Wi-Fi) received the internet and distributed to the surrounding devices. This AP can connected to different networks. The devices (mobile and computers) are able to access to the Internet through the AP. The proposed hybrid classifier (HC) was connected directed to the targeted AP. All the Internet traffic pass through this AP will be classified by our hybrid classifier.
The port-based classifier is part of the HC classification system which is entirely based on a port number. Basically, the classifier compares the coming flow port number with the saved ports. If the flow port number is similar to any of the saved port numbers, the flow will be classified based on the saved port. The approach is easy and fast and does not require sophisticated hardware. Furthermore, the port method checks only the transport layer header (TCP or UDP header) and does not check the other layers or the payload. The port table is a small database which includes the port numbers of the three classes which are considered by the HC experiments. Table 3 shows the class application port numbers which are used by port classifier. The rest of the port numbers that are not saved in HC port table will be classified by the port classifier as unknown.
Port classifier of HC system check the flow port number. If it is similar to any number that exists in the table, it will be classified equal to the class of that port number. Since the port number is numeric, the checking process is very fast and no delay is observed.

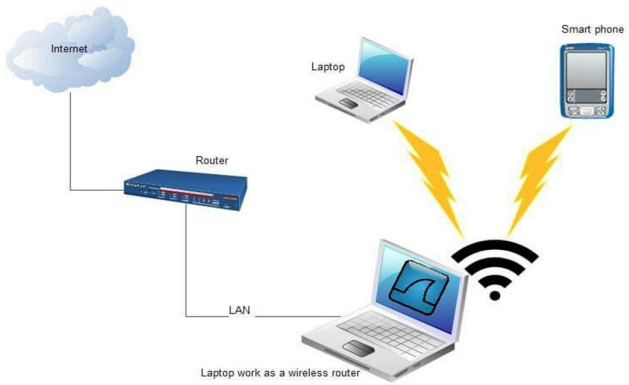

Figure 3 Network environment of proposed system

TABLE I: PORT NUMBERS FOR THE CONSIDERED APPLICATION CLASSES

Applications (class)	Port numbers
www (http, https)	80, 443,
WhatsApp	5228,5223
Twitter	-

The second part of HC is ML classifier which is define as statistical classifier. The statistical classifier plays an essential role in the HC system. Unlike port, the statistical classifier has a classification decision about the traffic flow in most cases. This means port classifier will ignore the flow if there are no well-known port numbers or the traffic is encrypted, whereas the statistical classifier will automatically make the classification decision without constraint. The statistical classifier in an HC system considers the following factors which can help improve the ML classification quality:

The statistical algorithm imports the statistical rules from the offline training stage and uses these rules to check to which class this traffic flow belongs. In the training stage, more than 10 Weka algorithms was tested. This includes class selection, features selection, algorithms selection, and building the classification rules. The statistical classifier in the HC system is a trade-off between the rules generated by one of the three algorithms i.e. rules.PART, Tree.J48, and RandomTree. The rules generated by the algorithm in the offline training stage are used for the classification (offline and online). After wide analysis the rules of rules.PART algorithm are used. As it is shown before, this algorithm generates rules which are less four times than the other algorithms. In addition, the accuracy gained by the rules of this algorithm is high.

The proposed hybrid classifier (HC) differs from others, since the classification decision is based on two different parallel hybrid methods. In addition, this classifier is not based on hardware component and does not make the classification based only on port-based method. In the proposed scenario, each of the two classifiers will individually classify the same traffic flow. Based on some priority rules, our HC makes classification decisions for each flow.

Table 1 shows the order of HC priority rules. The symbol (□) means the classifier has a decision about the flow traffic, whilst the symbol (×) means the classifier has no decision about this traffic flow or it classifies the flow as unknown. In the first rule, HC decision is “unknown”, because all the classifiers have no decision about the current traffic flow. In the second rule, HC classifies the flow as class A when both ML and port classifiers classify the flow to the same class (class A). In the third rule, the current flow is identified as class A by the port classifier and class B by the ML classifier. In this case, the HC decision is equal to the ML classifier (class B). In the last rule, HC classifies the flow as class A (based on port classifier) when ML classifier have no decision about this flow.

TABLE II: HC PRIORITY RULES ORDER

HC priorities order	Port classifier	ML classifier
1. Unknown	×	×
2. Port or Statistic classifier (class A)	✓ (class A)	✓ (class A)
3. Statistic classifier (class B)	✓ (class A)	✓ (class B)
4. Port classifier (class A)	✓ (class A)	×

IV. Validation and implementation Results

The proposed HC was tested by identify traffic of three types of applications WhatsApp, Twitter and http. All the datasets of these traffic was collected from campus environment (Umm Al-Qura University- Computing College at AlQunfudah). Wireshark software [20], was used capture and analysis these traffic. The captured file contains a large number of packets, carrying information and data about the captured application. During the capture process, manually only the needed traffic was generated and captured. This means, all the other Internet traffic was prevented from generated traffic. Even the windows and applications update was closed.
In ML classifier, Weka open source was used as machine learning tool in the training stage. The CSV file which prepared in the capturing step was used to prepare Weka file. Based on the previous steps, the rules of PART algorithm (Weka algorithm) was copied and saved. This rules was prepared to be used by MATLAB which are involved in if else statement.
Table 2 illustrates number of packets which used in training and testing stage for each of the considered application.

Table III. Number of packets in training and testing stages

Application	Number of training packets	Number of testing packets
http	1500	750
WhatsApp	1500	750
Twitter	1500	750

As mentioned before the hybrid classifier (HC) makes his decision based on the decision of both ML and port classifier. .
Rules generated by PART algorithm were used by HC. Figure 5 illustrate the results of Twitter traffic classification. The figure shows the accuracy of the proposed HC compared to ML and port classifier. As seen the HC and ML classifier provide a high accuracy (72.5%) compared to port classifier (0%).

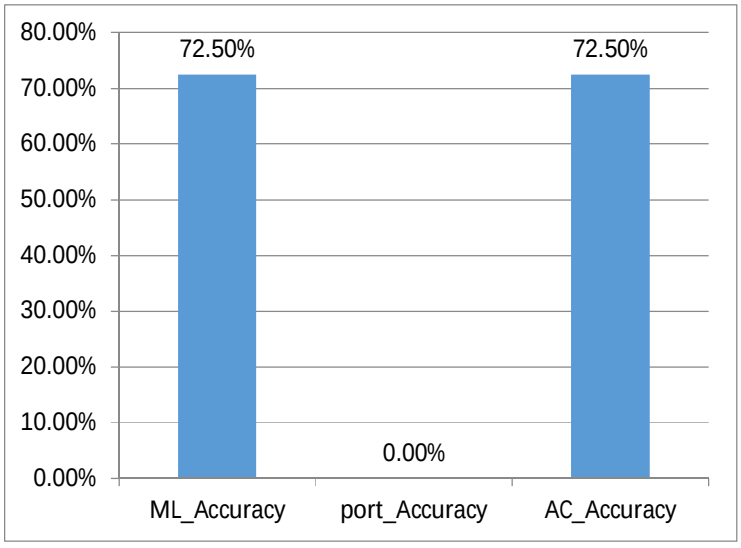

Figure 4 Twitter traffic classification results

The testing data of http traffic was classified by the hybrid classifier. Figure 5 shows the results of http traffic classification. As shown in the figure, hybrid classifier has a high classification accuracy (90.13%) compared to ML (84.8%) and port classifier (55.73%).
In the same way, the testing data of the WhatsApp traffic was classified by the proposed HC. Figure 6 shows WhatsApp traffic classification results. As it's shown, the hybrid classifier generate a high accuracy (88.79%) when it compared with the other two classifier (85.05% and 37.9%).

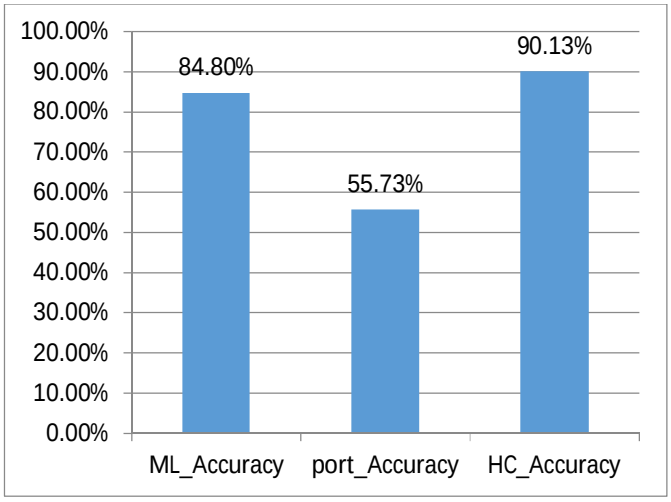

Figure 5 http traffic classification results

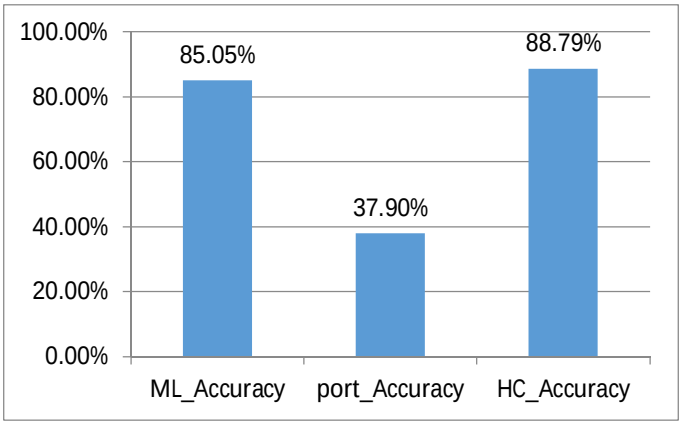

Figure 6 WhatsApp traffic classification results

V. Conclusion

The classifier that uses only one method is constrained by the limitations of that method. Although, machine learning approach is appropriate in identifying Internet traffic and resolves the problems of classifying unknown port and encrypted traffic, however, this approach still has some limitations such as features overlapping and ML dataset scenarios. The combination of more than one method is aimed to utilize all the benefits of the individual classifiers in only one main classifier. As an example, the statistical-based classification method (ML method) has the ability to classify encrypted traffic, and the port-based classification method has the advantage of simplicity. Combining these two methods will result in a classifier that is simple and is able to identify encrypted traffic at the same time.

This paper proposed a hybrid classifier to classify internet traffic based on two classification methods port and statistical methods. The goal is identify each packet of the Internet traffic based on their application type. The proposed hybrid classifier was tested by classifying three types of internet applications (http, WhatsApp, and twitter). Wireshark was used to capture real network traffic which is analyzed and filtered in manual stage. This captured file was prepared for Weka by adding Weka header and data. More than ten of Weka algorithm was applied to train the ML classifier. The output of the training stage is statistical rules was used in the hybrid classifier. WhatsApp, http, and twitter traffic was identifying using the proposed classifier. The classification shows accepted results for each of considered internet application.

REFERENCES

[1] Callado, A., et al., A survey on internet traffic identification. Ieee Communications Surveys and Tutorials, 2009. 11(3): p. 37-52.

[2] Bujlow, T., T. Riaz, and J.M. Pedersen. A method for classification of network traffic based on C5.0 Machine Learning Algorithm. in Computing, Networking and Communications (ICNC), 2012 International Conference on. 2012.

[3] Yamansavascilar, B., et al. Application identification via network traffic classification. in 2017 International Conference on Computing, Networking and Communications (ICNC). 2017. IEEE.

[4] Kim, J., J. Hwang, and K. Kim, High-performance internet traffic classification using a Markov model and Kullback-Leibler divergence. Mobile Information Systems, 2016. 2016.

[5] Ye, W. and K. Cho, P2P and P2P botnet traffic classification in two stages. Soft Computing, 2017. 21(5): p. 1315-1326.

[6] Sun, M.F. and J.T. Chen, Research of the traffic characteristics for the real time online traffic classification. Journal of China Universities of Posts and Telecommunications, 2011. 18(3): p. 92-98.

[7] Min, D., C. Xingshu, and T. Jun, Online Internet traffic identification algorithm based on multistage classifier. Communications, China, 2013. 10(2): p. 89-97.

[8] Molnar, S. and M. Perenyi, On the identification and analysis of Skype traffic. International Journal of Communication Systems, 2011. 24(1): p. 94-117.

[9] Yu, J., et al., Real-time Classification of Internet Application Traffic using a Hierarchical Multi-class SVM. Ksii Transactions on Internet and Information Systems, 2010. 4(5): p. 859-876.

[10] Soysal, M. and E.G. Schmidt, Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. Performance Evaluation, 2010. 67(6): p. 451-467.

[11] Gu, R., H. Wang, and Y. Ji. Early traffic identification using Bayesian networks. 2010.

[12] Nguyen, T.T.T., et al., Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic. Networking, IEEE/ACM Transactions on, 2012. PP(99): p. 1-1.

[13] Chen, Z.X., et al., Online hybrid traffic classifier for Peer-to-Peer systems based on network processors. Applied Soft Computing, 2009. 9(2): p. 685-694.

[14] Singh, K. and S. Agrawal. Comparative analysis of five machine learning algorithms for IP traffic classification. in Emerging Trends in Networks and Computer Communications (ETNCC), 2011 International Conference on. 2011. IEEE.

[15] Ruoyu, W., L. Zhen, and Z. Ling, A New Re-sampling Method for Network Traffic Classification Using SML. 2010.

[16] Ye, W. and K. Cho, Hybrid P2P traffic classification with heuristic rules and machine learning. Soft Computing, 2014. 18(9): p. 1815-1827.

[17] Adami, D., et al., Skype-Hunter: A real-time system for the detection and classification of Skype traffic. INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS, 2012. 25: p. 386–403.

[18] Aggarwal, R. and N. Singh, A new hybrid approach for network traffic classification using SVM and Naïve Bayes algorithm. Int. J. Comput. Sci. Mobile Comput, 2017. 6: p. 168-174.

[19] Aouini, Z., et al. Early classification of residential networks traffic using c5. 0 machine learning algorithm. in 2018 Wireless Days (WD). 2018. IEEE.

[20] Orebaugh, A., et al., Wireshark & Ethereal Network Protocol Analyzer Toolkit. 2007: Syngress.