International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Nov-2021)
http://dx.doi.org/10.12785/ijcds/1001117

Ensemble Machine Learning for P2P Traffic Identification

Md. Sarfaraj Alam Ansari[1], Kunwar Pal[2], Mahesh Chandra Govil[1], Prajjval Govil[3] and Adarsh Srivastava[1]

[1] Department of Computer Science and Engineering, National Institute of Technology Sikkim, India
[2] Department of Computer Science and Engineering, Dr B R Ambedkar National Institute of Technology Jalandhar, India
3Department of Computer Science and Engineering, JK Lakshmipat University, Jaipur, Rajasthan 302 026, India

Received 15 Nov. 2020, Revised 26 Mar. 2021, Accepted 03 Apr. 2021, Published 25 Nov. 2021

Abstract: Network traffic identification and classification in the current scenario are not only required for traffic management but in designing a future protocol for user-specific services and improve user experiences. This fundamental step of network management is perceived by the researcher long back and started developing techniques for the same. The traditional techniques for traffic identification and classification include port and payload based. The current large and complex network poses many challenges to the researchers in designing approaches for traffic classification by using dynamic ports, encryption, and masquerading techniques. The complexity is further enhanced due to increased dependence on the Internet and diverse applications to enable network administrators including ISPs to manage the network intelligently and efficiently. As traditional techniques are not effective to address the current challenges, a hybrid solution is explored. The hybrid approaches make use of statistics or behavioral-based, heuristic-based, machine learning-based along with feature selection techniques. In this paper, apart from developing enhanced hybrid approaches for identifying the P2P traffic, an extensive real dataset of size 924 GB is constructed to analyze the effectiveness of the proposed approaches. A number of hybrid approaches are designed by using feature selection techniques and machine learning (ML) algorithms. Extensive analysis of proposed hybrid approaches along with the comparative study reveals that Chi-Square and Random Forest outperform other state-of-art approaches yielding an accuracy rate of 99.46%.

Keywords: Internet Traffic, Peer-to-Peer (P2P), Feature Selection, Classifier, Machine Learning

1. INTRODUCTION

Internet, since its evolution, has witnessed tremendous growth due to ease of various online services and so in content distribution and sharing. The popularity of the Internet can be attributed to the untiring efforts of researchers in paving the path for unprecedented information technology developments in almost all spheres of communication especially, hardware and software systems. In the past, client-server is the commonly used model for content distribution in which the server delivers the requested contents to the client. The traditional client-server system has serious limitations to cope with the requirements of modern large, dynamic, and complex computer networks. At present, the usage of Internet is increasing many folds and the Internet has become an indispensable platform in almost all fields of life like entertainment, education, business, etc. The dramatic growth in online applications and the need for certain basic quality of service (QoS) requirements such as scalability, delay in content delivery, resource usage, user experience, bandwidth, etc. have made client-server systems inadequate for the modern demand of massive, dynamic, and diverse Internet traffic.

The evolution of peer-to-peer (P2P) network is one of the probable solutions for the problem. In the P2P network, each of the peer serves both as client and server at the same time which makes the system distinguishable from conventional client-server architectures [1]. Further, the network architecture should also be able to deal with network congestion, robustness, scalability, cost-effectiveness, the fulfillment of user expectations, QoS requirements, traffic hindrance, and efficient use of bandwidth [2]. Resource sharing capability of P2P systems is an important property that allows all individual devices and multiple peers to harness the unified power to get benefits [3]. Again, in P2P networks, traffic is symmetric and increment in user numbers rarely leads to network performance degradation unlike in the client-server system [4] where the structure is inherently non-sharing and network traffic is asymmetric due to unidirectional content delivery structure.

The exponential growth of P2P traffic can be attributed to the dominance of P2P applications over the Internet in comparison to other applications viz FTP,

E-mail: sarfaraj@nitsikkim.ac.in, kunwarp@nitj.ac.in, govilmc@gmail.com, prajjvalgovil@gmail.com, b180001@nitsikkim.ac.in

HTTP, SMTP, etc. P2P applications not only include video, audio, and gaming that contributes to a huge data transfer but in recent years, the P2P file-sharing trend has also added to the size of data sharing and distribution. It is reported that a major portion of the Internet traffic is Peer-to-peer (P2P) and is still growing [5]. It occupied nearly 70% of the Internet traffic and hence consumes a major portion of Bandwidth [6]. Another study [7] has revealed that the Asymmetric Wireless Subscriber Line (ADSL) traffic is 49 percent due to P2P applications. As per CISCO VNI 2020 forecast, globally, video traffic on the Internet will rise four times between 2015 and 2020, with an annual growth rate of 31 percent [55]. The wise and fair utilization of network resources is one issue, but research must provide answers to myriad questions to achieve ultimate goals to satisfy user's needs and expectations. Thus, current large-scale P2P applications have brought a serious need for monitoring and controlling network traffic. In the P2P network, among other performance issues, chunk scheduling [8] [9], flash crowd [10] [11] [56], selfish peers [12] [13] are major concern apart from traffic classification. Selfish peer in the P2P network is considered one of the major problems which can severely degrade the performance and is negation to the basic architectural nature of P2P networks. This kind of peer's behavior is also termed as free riding, as the peers consume the resources without sharing in return. What is desirous is proper network management and intelligent traffic analysis techniques. In our opinion, identification and categorization of network traffic crossing the network boundaries is the first step to enable the network administrators to implement necessary fine-grained traffic management and the policies for security [6]. The other major concern which imposes the crucial need of traffic identification and categorization are ISPs challenges such as paying for additional traffic requirements, excellent customer satisfaction, cost of bandwidth, implementing billing mechanism, implementation80 of application-specific policies; maintaining QoS of applications; implementing security measures; etc. Further, the task of traffic classification is considered as future solutions for addressing various P2P network problems, new protocol design, developing methods for network security to handle attack detection and prevention, flow cleaning, etc. [14].

As traffic identification and classification provide a sound platform for network management effectively. The journey of this research can be traced back to traditional methods relying on well-known service port numbers [15] [16] [59]. It is popular because it’s simplicity, ease of implementation, and does not involve much calculations. For example, DNS or SMTP uses specific ports statically, therefore, yields high accuracy of classification. As the years progress, the use of random port numbers and masquerading technique across the applications have become common making port-based classification inefficient [17] [18]. Further, the encryption techniques are aggravated the traffic identification task [19]. Payload based identifications [20] is another mainstream approach used and rely on deep packet inspection. Though this technique yields higher accuracy, it suffers from high computational overhead, user privacy issues, and very low accuracy when data is encrypted [21]. In addition, owing to its complexity and processing burden on network equipment, it is impractical for high-speed networks. Park et al. [22] have highlighted an important fact that although port-based methods provide low classification accuracy, this method is still relevant in the Internet backbone due to its scalability and minimal computational overheads. Hence, port-based approaches play a determining role to give a direction when combined with other methods to make a hybrid approach for identifying the P2P traffic. In the past few years, researchers are giving more emphasis on exploring other approaches such as statistics or behavior-based, heuristics-based, and machine learning-based to identify and classify Internet traffic. It is observed that each technique has its own limitations. The applications which have similar behavior are difficult to analyze with the behavioral based approach. In the case of the statistical-based method, the numerical attributes do not always provide high-quality training data. The ML techniques are intelligent and flexible, but they are also facing many challenges such as optimal feature selection, high dimensionality issues, and high correlation between traffic classification accuracy and the prior probability of training data [23] [57]. It is projected that the integration of different native techniques will provide the desired accuracy and QoS and hence, the research has moved to the development of hybrid approaches. There are several reasons to look for ML-based methods to define and classify P2P traffic. Besides efficient network management, there are also many other issues such as selfish peers, flash crowds etc. need to be addressed.

The paper presents ML-based techniques using network attributes to identify P2P traffic. The main purpose of this work was to construct a dataset which can be used for the study of various P2P network issues to cater the need of customization of services in modern network. In this paper, hybrid approaches are developed for classifying Internet traffic into P2P and non-P2P by leveraging the advantages of various methods mentioned above. The proposed approaches are an amalgamation of the port-based method, Feature Selection (FS) techniques, and ML Algorithms. The salient contributions of this research work are as follows:

- Construction of 'SAMPARK' dataset of size approximately 924 GB by employing techniques for data collection in line with the literature.
- Pre-processing of collected data and feature extraction.
- Study of the impact of feature selection techniques and their applications.
- Analysis of the effectiveness of five ML algorithms.
- Quantitative analysis of different hybrid approaches developed by combining five ML algorithms (Random Forest (RF), Decision Tree (DT), K-Neural Network (KNN) Naive Bayes (NB), and Support Vector Machine (SVM)) with feature selection methods (Chi-Square (χ^2), Analysis of Variance (ANOVA) and Principal Component Analysis (PCA)) on SAMPARK and UNIBS sub-datasets.

The organization of this paper are: section 2 covers a detail of related work; the proposed methodology which includes construction of dataset, and proposed approaches for Internet traffic classification are presented in Section 3. Section 4 covers the experimental setup and performance analysis. Lastly, section 5 draws the final remarks and possible future work.

2. RELATED WORK

The port number-based technique [15] [16] as discussed earlier is simple to use and implement. However, the purely port-based traffic classification techniques have become nearly ineffective as the increased usage of dynamic port number, masquerading, and encryption techniques [19]. The traffic identification is limited to those applications that have known port numbers with certainty although the accuracy is very high for such applications. Jeffrey et al. [24] and Bhatia et al. [59] have advocated that port-based techniques are still useful and can provide better results. Similarly, the payload-based techniques are no longer efficient in their intrinsic form due to the reasons mentioned earlier. The payload-based technique can detect the traffic for which signatures are known but fail to classify unknown traffic.

The present trend in the research community is to design and develop hybrid approaches that combine various techniques from different domains such as statistical or behavior-based, heuristic-based, Machine Learning [25] [26] [27] [28], Genetic Algorithm, and Neural Network, etc. [29] with intrinsic methods. The approaches which are independent of port number and payload inspection can be grouped under classification in the Dark [21] [30].

Statistics or behavioral-based approaches identify Internet traffic according to statistical features collectively or independently [58] [59]. The example of statistical features are flow size, flow count, size of first packet in flow, inter-arrival time of packet (Pkt_IAT), flow duration, etc. These can be extracted from the traces. It is believed to have each traffic class generated by different applications have unique characteristics. Considering the approach, various works have been proposed by using different feature combinations as discussed in the literature [31] [32] [33] [34]. But it becomes difficult to map between the increasing number of characteristics against the corresponding traffic class and therefore need to combine with other methods to yield results such as heuristics or machine learning. The heuristics appear to be promising solutions to identify and classify network traffic and notable work [20] [25] [35] [36] is done by research in this area also. The packet and flow-level behavior details of traffic are explored to develop novel methods. The benefit of such strategies is its ability to generalize the learned activity to work well with unknown applications and thereby increases the ability to track class of Internet traffic. The pre-defined traffic pattern such as the packets sent/received by a peer, a peer connected with distinct number of hosts, the total number of connections made by a host, upload-download ratio, etc. is some of the traffic/application characteristics which are used to synthesize the heuristic functions. Machine learning algorithms are increasingly used in almost all domains due to their intelligent and flexible nature as discussed earlier. The ML techniques have demonstrated good performance in traffic classification also. In recent years, the shift in focus to develop approaches using ML techniques has been noticed [14]. By integrating different techniques discussed above, the paper reported using approaches focused on both ML methods and hybrid approaches.

Jeffrey et al. [24], in their proposed work, have used unsupervised machine learning approached and compared the result with a previous supervised machine learning-based approach. They have demonstrated that the proposed unsupervised classifier outperformed the supervised classifier by 9% on 1000 samples of each class. Raahemi et al. [26] have used the CVFDT technique and obtained 95% accuracy. They have collected their own dataset and determined the performance for every 10,000 examples. In [29], traffic identification is based on the genetic algorithm and neural network. The accuracy claimed is nearly 96% on their own dataset consisting of 32767 sample records. In [24] [26] [29], the dataset is labelled using the default port numbers of P2P applications. Hussein et al. [37] have achieved up to 98% accuracy on the dataset collected by accessing BBC, Facebook, Google search, Skype, Yahoo Mail, and YouTube separately executing each application 30 times for 2-5 minutes. It analyses the timing features of the burstiness of induced traffic against each application and used the C5.0 classifier for network traffic characterization. In [25] and [37], have investigated the behavior-based features to train and test the data.

Draper-Gil et al. [27] have also explored time-based features for capturing the VPN traces. KNN and C4.5 classifiers have been used to classify the extracted features into different categories and concluded that time-related features can be a good choice in the identification of network traffic. The accuracy obtained is approximately 80%. Further, Saber et al. [28] have tried to enhance the above approach [27] using PCA for feature selection and classify the combined over & under-sampled data of VPN and non-VPN using SVM. They have reported accuracy of 96.6%, 95.6%, 93.9%, 94.9% while considering the flow time-outs of 15s, 30s, 60s, and 120s respectively. However, the efficiency is higher for shorter flows. It demonstrates that the application of feature selection techniques can give better traffic classification efficiency. Junior et al. [38] have used ANOVA as a feature selection technique with some clustering algorithms and achieved P2P classification accuracy of 90%.

Bhattacharya et al. [39] used KNN, NB, RF, SVM, and XGBoost classifiers in combination with PCA and hybrid PCA-firefly algorithms foe performance evaluation while classifying the Intrusion Detection System (IDS) data set. The proposed hybrid PCA-firefly with the XGBoost model is found to achieve more than 99% accuracy. They have used Kaggle dataset of 125973 instances and performed the experiments on the Google Colab GPU platform. Wang et al. [40] have developed a network traffic identification method using SVM and achieved 99.31% accuracy with regular biased training and test data set. They have exported the traffic from the network using MATLAB and LibSVM. In [23], an improved model using SVM classification is developed for network traffic classification and achieved the accuracy of 99.34%. It is provided better accuracy compared to KNN, NB, RBFNetwork, and SVM with 24897 samples and claimed to perform better. The dataset used is limited in size and collected fifteen years ago.

3. Methodology

In this paper, hybrid approaches are explored for classification of Internet traffic into P2P and non-P2P. Novel approaches are developed as an amalgamation of port-based methods, feature selection techniques, and ML Algorithms.

A. Framework

The objective of the proposed work is to provide an efficient ML-based identification model for P2P application from Internet traffic. The working procedure of the proposed methodology is demonstrated in Figure 1. There is a requirement for a big, real-time, and trusted dataset for verification of an efficient model. So, a real dataset (924 GB) is collected from the Internet using Wireshark. The first sub-module of the given framework is the collection of data and it’s pre-processing which represented by the rectangular boxes in Figure 1. After pre-processing, the major task which needs to be performed is feature selection. But before that proper labelling needs to be performed. For the labelling of the dataset, a list of port numbers is used. The list is extracted from the dataset, also gathered from the literature, and further, it is used to label the dataset for training set purposes. Next sub-module is feature selection, which extract the selected features applying different feature selection techniques and after that selected feature are evaluated by using different ML algorithm for classifying P2P and non-P2P traffic. Further decision making on different combinational modules is performed using different performance matrix.

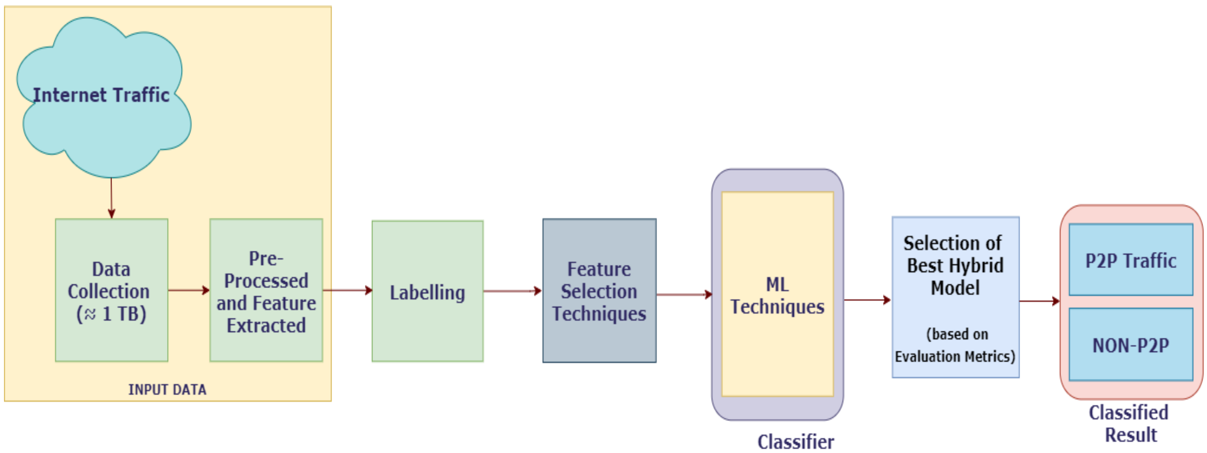

Figure 1. Framework for proposed methodolog

Finally, an efficient and more accurate model is achieved by amalgamation of feature selection and ML approaches for more accurate identification of P2P and Non-P2P traffic. Basically, the framework indicates proposed identification method, which is supposed to be as automated as possible, based on that it can achieve comparable accuracy. Therefore, the methodology contains data collection, pre-processing, feature extraction, labelling of the dataset, feature selection, and ML Algorithm, and these are presented systematically in the framework.

B. Data Collection

The real dataset is necessary to test any proposed ML-based approach comprehensively. If the dataset is available publicly, it saves the information gathering time and improves the research productivity. But most of the existing dataset is developed a decade ago, suggesting that it is different due to the network current speed and rapid application revolution [41]. Also, due to privacy and security reason, the availability of labelled data sets is very limited [42]. Therefore, it is better to carry out research work by collecting real-time data. Extensive efforts have been put in to collect the dataset by capturing the traces from varied applications in the network. A real dataset named SAMPARK has been constructed to address the issues of P2P networks in this work presented. The Wireshark application is the most beneficial and free software. The collection of data from Internet traffic using Wireshark is very effective from the research perspective. Wireshark gives the data into PcapNG file then it has to be converted the data into CSV file for ease of computation. For collection of traces, the National Institute of Technology Sikkim ICT infrastructure is used. Institute provides us more than 10 public IP addresses to collect the traces. Private IP addresses are being used when the traces are captured from the peers at the Computer Network Laboratory of the Institute. The testbed for data collection is presented in Figure 2. The traces are collected and stored batch-wise, each batch contains five cases and multiple numbers of peers used in each case to collect the traces. It is difficult to handle a huge dataset. Therefore, raw Internet traces are captured for an hour basis by- running multiple applications. The total raw data size of 924 GB is obtained during 2nd - 9th October 2019. A name 'SAMPARK' for the raw captured dataset is proposed for ease of communication. Wireshark provided the data in PcapNG files. The collected dataset is converted into CSV files for ease of the feature extraction process. The traces are also collected for 10 minutes by running individual applications such as BitTorrent, PPTV, Funshion, Vuze, Miro, Skype, QQplayer, µTorrent, Tribler, YouTube, iQIYI through the Internet on each peer so that ports can be extracted for labelling the dataset. This sample dataset is named as 'MKS' dataset. The detailed analysis of the 'SAMPARK' dataset is mentioned in TABLE I. The protocol-wise flow details are presented in TABLE II. 'SAMPARK' dataset is explored in this paper. The research communities who are practicing similar research will be benefitted from this dataset. The dataset is collected in such a way that it helps to find free riders, malicious peers, etc.

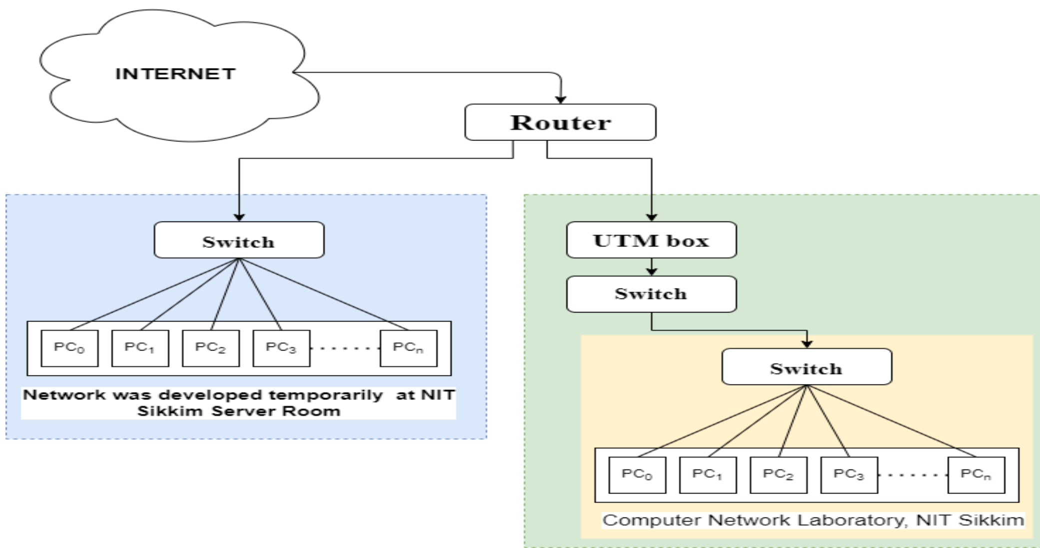

Figure 2. Testbed for data collection

To validate our work, a popular dataset 'UNIBS' [43] [44] is considered for the necessary comparison. The UNIBS dataset is used by the research community for similar work. The UNIBS dataset includes packets generated in University of Brescia, Italy in 2009. Tcpdump is used to capture these traces which include the classes, such as Mail, Web, SKYPE, P2P, etc. TABLE III shown details of UNIBS traces.

C. Pre-processing of Data and Feature Extraction

The data pre-processing and extraction of different features subsets are the important tasks in a hybrid approach. These are massive tasks and if not address adequately may affect the results. Pre-processing of data is an important task as it filters input data to create a set of patterns by removing identical, extraneous, and(or) noisy features to minimize the errors in the results. Packet level and flow levels have different network traffic characteristics [45]. Many packets together constitute a flow, characterised by same source and destination IP address, protocol, source, and destination port. The flow-based features are used for the classification of internet traffic and identification of P2P applications. The packet-based features (e.g., Up/Down Ratio) are to be considered for various purposes, such as identifying selfish peers in the network. Network connections have parallel bidirectional communication between two peers. The uplink/downlink packets are decided according to the first packet of the flow. Most of the features are self-explanatory and briefly described in the TABLE IV. However, some flow-based features are defined below for a better understanding of the discussion.

Definition 1 (Flow duration): Let F_d be denoted as flow duration then,

$$F_d = t_n - t_1 \qquad (1)$$

Here, t_1 and t_n are the time stamp of the first and last packets in the flow.

Definition 2 (Flow count): The flow count, denoted as F_c, is defined as the total packet numbers in a flow.

Definition 3 (Flow Size): Let F_s be denoted as flow size, then

$$F_s = \sum_{s=1}^{n} P_s \qquad (2)$$

where P_s be each packet size and n be packet numbers in the flow.

Definition 4 (Packet Inter Arrival Time): It is time interval from ith packet to (i−1)th packet arrival time and denoted as Pkt_IAT. The figure 3 shown a glimpse of python code written for calculation of packet inter arrival time.

D. Features Selection Algorithm

In general, machine learning operates on large and concise datasets. But using high dimensional data has various pitfalls among which the major one is the curse of dimensionality [46].

As a result, computation time is increased, and data pre-processing and EDA (Exploratory Data Analysis) are more complicated. These are due to the redundant features available in the dataset and inconsistencies present in the features. The problem discussed above can be solved by a process or method called reduction of dimensionality. It is a method used to filter out important features necessary for the purpose of training. There are various algorithms available for the purpose, however, in this paper χ^2, ANOVA and PCA are considered as feature selection techniques. The detailed discussion is as follows.

1) Chi Square Test

It is used to assess the discrepancy in design or by some significant factor in the predicted value and observed value if any. The value of chi-square (χ2) is calculated by the formula mentioned in equation no. (3). We are taking summation of squared value of difference of observed and expected value divided by the expected value. The degrees of freedom of the (χ2) are calculated as one less than the number of observations.

A pre-defined chi-squared distribution table is there from which the critical chi-squared value can be obtained against the degree of freedom and significance level. Now, compare it with the critical chi-square value. If the obtained value or percentage is low, it indicates the high correlation of two features in the dataset. The performance of the Chi-square test is very effective since it has the ability to perform well as a method of feature selection [47]. The (χ2) value of attribution is shown below:

$$\chi^2 = \frac{(F_0 - F_e)^2}{F_e} \quad (3)$$

where the observed value is denoted by F_0, F_e represents the expected value. A rank can be determined for each feature based on (χ^2) value of all listed features. The features with high rank are given more priority than the other [48] [49].

2) ANOVA

We can effectively analyze the complex data by finding the statistically relevant mean difference between the groups using one-way ANOVA (Abdalla et al. 2017). Here, the F-ratio and degree of freedom are calculated. A predefined table is available against the significance level of 0.05 and 0.01. The table is based on the degree of freedom of "Variance difference between the groups" and "Variance difference within the groups".

TABLE I. DETAILS OF SAMPARK AND MKS DATASETS

Batch	Case	Nos. of PcapNG files	Total Data Size (in GB)
SAMPARK Dataset (Raw Internet Traffic)			
Batch 1	Case 1	06	26.2
	Case 2	06	20.7
	Case 3	06	21.0
	Case 4	08	53.2
	Case 5	08	41.9
Batch 2	Case 1	08	13.3
	Case 2	08	53.8
	Case 3	08	66.8
	Case 4	08	67.0
	Case 5	08	64.5
Batch 3	Case 1	08	42.3
	Case 2	08	43.0
	Case 3	08	94.0
	Case 4	08	48.8
	Case 5	08	37.9
Batch 4	Case 1	10	22.5
	Case 2	10	30.9
	Case 3	10	20.4
	Case 4	10	25.1
	Case 5	09	10.8
Batch 5	Case 1	10	7.37
	Case 2	10	17.6
	Case 3	10	26.0
	Case 4	10	15.0
	Case 5	10	21.9
MKS Dataset (Application based)			
10 min Data	14 application	28	6.97
1 hr Data	10 application	10	23.9
Total counts:		241	924

TABLE II. PROTOCOL-WISE FLOW DETAILS OF SAMPARK DATASETS

Sl. No.	Protocol	Count of Flow	% of flow
0	UDP	3561898	51.01
1	TCP	3037260	43.50
2	ICMP	104640	1.50
3	BitTorrent	77844	1.11
4	DNPv100	49762	0.71
5	HTTP	27857	0.40
6	TLSv1.2	23542	0.34
7	ICMPv6	15526	0.22
8	DNPv65	12786	0.183
9	DNPv17	10353	0.148
10	DNPv33	9839	0.141
11	DPPv0	7099	0.102
12	TLSv1.3	4278	0.061
13	ARP	4245	0.061
14	HTTP/XML	3403	0.049
...	...	...	...
...	...	...	...
Total counts			
486	...	6982319	100.0

TABLE III. DETAILS OF UNIBS DATASETS

Dataset	Data Size
unibs20090930.anon	317 MB
unibs20091001.anon	236 MB
unibs20091002.anon	1.94 B

If the F-ratio is more than the significance level of 0.05 and 0.01 values, then we reject the Null Hypothesis otherwise we accept it. For this reason, Sum of square (SS), Sum of Squares for Treatment (SST), Sum of Squares for Error (SSE), Variance Between Treatments (MST), Variance Within Treatments (MSE) are computed and represented mathematically by the following equations no. (4) through (13):

$$SS = \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{i,j} - \bar{y})^2 \quad (4)$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{i,j} - \bar{y} + \bar{y}_j - \bar{y})^2 \quad (5)$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{i,j} - \bar{y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_i}(\bar{y}_j - \bar{y})^2 \quad (6)$$

$$SS = SST + SSE \quad (7)$$

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{i,j} - \bar{y})^2 \quad (8)$$

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_i}(\bar{y}_j - \bar{y})^2 \quad (9)$$

$$SS = SST + SSE \quad (10)$$

$$MST = \frac{SST}{k-1} \quad (11)$$

$$MSE = \frac{SSE}{k-1} \quad (12)$$

TABLE IV. LIST OF EXTRACTED TRAFFIC FEATURES FROM BOTH SAMPARK AND UNIBS DATASETS

Feature No.	Feature Name	Description
1.	Src_ip	Source IP address
2.	Dst_ip	Destination IP address
3.	Protocol	Transaction protocol (TCP, UDP, etc.)
4.	Src_port	Source Port Address
5.	Dst_Port	Destination Port Address
6.	Flow_count	Nos. of packets appeared in a particular flow
7.	Flow_size	Total sent or received data by a particular flow
8.	Pkt_size_of_first_flow	Size of a packet when it appears first in a flow
9.	Flow_duration	Total flow duration
10.	Flow_IAT	Inter arrival time of flows
11.	Pkt_IAT_as_source	Inter packet arrival time as source
12.	Nos._of_times_as_source	Nos. of times it appears as source
13.	Mean_sq_pkt_size_as_source	Mean square of packet size transmitted by the source
14.	Data_of_first_pkt_as_source	Total bytes in a packet when a IP appears first at source
15.	Total_data_sent_as_source	Total data sent by an IP when it appears as source
16.	Control_byte_sent_as_source	Total bytes sent by a control packet when the IP appears as source
17.	Pkt_IAT_as_destination	Inter packet arrival time as destination
18.	Nos._of_times_as_destination	Nos. of time it appears as destination
19.	Mean_sq_pkt_size_as_destination	Mean square of packet size received by the destination
20.	Data_of_first_pkt_as_destination	Total bytes in a packet when a IP appears first as receiver at destination
21.	Total_data_recv/send_as_destination	Total data received by an IP when it appears as destination
22.	Control_byte_sent/recv_as_destination	Total bytes received by a control packet when the IP appears as destination
23.	Total_duration	Total duration of an IP which participated in the network
24.	Ratio_up/down	The ratio of total bytes sent and received by an IP

$$F = \frac{MST}{MSE} \tag{13}$$

Test H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ and H_1: $\mu_1 = \mu_2 = \cdots = \mu_k$ where $\bar{y}$, n and y_j denote the samples mean, sample size and specified population mean, respectively. We considered ANOVA for feature subset selection. In order to improve predictive accuracy and prevent incomprehensibility due to the high number of features explored. To predict more accurate, different feature subsets have been considered and compared to get a competent model.

3) *Principal Component Analysis (PCA)*

PCA [50] helps us to figure out the correlation and patterns in the datasets. Due to this, a new dataset with a significantly lower dimension is formed, and most importantly during the process, there is no loss of any important information. The new variables that are derived from the initial sets are termed as Principal Components. The variables thus created are independent of one another and are highly significant. PCA algorithms comprise of following steps:

- Standardisation of the dataset: It means scaling the data in such a manner that all the variable and their values lie within the similar range.

$$Z = \frac{Variable\_Value - Mean\ (\mu)}{Standard\_Deviation\ (\sigma)} \tag{14}$$

- Covariance Matrix computation: It expresses the correlation between variables by maintaining the dependencies. Let Cov_{mat} be a covariance matrix of m x m dimensions. Let x and y be the features. Then,

$$Covariance(x, y) = \sum \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{N-1} \tag{15}$$

$$Cov_{mat} = \begin{pmatrix} cov(x,x) & cov(x,y) \\ cov\ (y,x) & cov\ (y,y) \end{pmatrix} \tag{16}$$

  where N=Number of elements; $\bar{x}$, $\bar{y}$ are means of x, y values. The (-) ve value of covariance indicates that the variables are indirectly proportional to each other whereas (+) ve value indicates that the variables are directly proportional to each other.

- Calculate the Eigenvector and Eigenvalues: The dimension of the dataset represents the eigenvectors to be calculated. Eigenvectors are used as the variance matrix to detect the data where most variances are there. The maximum variance in the data indicates more information. Principal Components are calculated using these eigenvectors.

$$Cov_{mat} - \lambda * Identity\ matrix = 0 \tag{17}$$

- We solve the equation and get two values of λ by putting the first value of λ in matrix and form an equation with x_1 and y_1 like AX =λX. Similarly, the second value of λ and do the same. We will get a matrix by these calculated values called as eigenvector matrix.
- Computing Principal Component: The higher value of λ and its subsequent eigenvector will be principal component. The less significant Principal Components are removed to reduce dimensions.
- Reducing the dimension of dataset: Rearrange the original data according to the most significant Principal Component.

E. *Machine Learning Algorithms*

Recently, machine learning algorithms have received significant attention in various fields. Researchers are also focusing more on ML based network traffic classification. Several classification algorithms are there to classify the traffic in P2P and non-P2P. In this work, various classification techniques such as DT, RF, NB, KNN, SVM are investigated, and the comparative analysis is presented in section 4.

TABLE V. PORT NUMBERS USED BY POPULAR P2P APPLICATIONS

P2P Applications	Port Numbers
BitTorrent	6881-6889
Edonkey (eMule, xMule)	2323, 3306, 4242, 4500, 4501, 4661-4674, 4677, 4678, 4711, 4712, 7778
Gnutella	6346, 6347
FastTrack	1214, 1215, 1331, 1337, 1683, 4329
DirectConnect (DC++)	411, 412, 1364-1383, 4702, 4703, 4662
Napster (File Navigator,WinMx)	5555, 6666, 6677, 6688, 6699-6701, 6257
Freenet	19114, 8081
Blubster	41170-41350
GoBoogy	5335
HotLine	5500-5503
ICQ	5190
IRC	7000, 7514, 6667
XMPP	5222, 5269
SoulSeek	2234, 5534
QNext	5235-5237

F. *Port Analysis and Labelling of Dataset*

In the proposed approaches, the port-based labelling of training data is carried out. As revealed in literature, the port-based approaches are better suited for data labelling due to ease of implementation and achieve higher accuracy. The process of port analysis and the labelling of dataset is presented graphically in Figure 3. For labelling the datasets, a list of known P2P port numbers is prepared considering both source and destination port. The list includes well-known, registered, ephemeral ports. The list prepared from our own collected MKS dataset is further expanded by including the port numbers gathered from the literature of a similar domain [3] [20] [29] [51] [52] [60]. The list of port numbers prepared by us is very large as compared to the list used by other researchers. The extracted port numbers in the prepared list are more than 22000 and are difficult to report here. Therefore, a glimpse of the list of port numbers is given in TABLE V and TABLE VI.

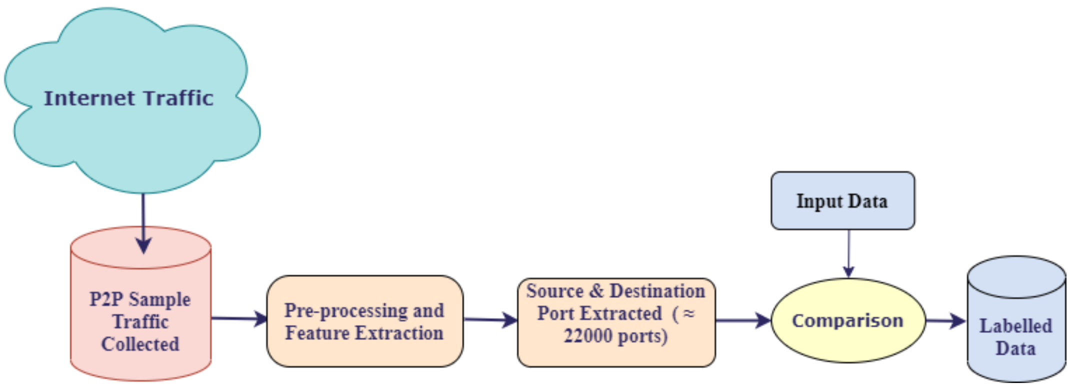

Figure 3. Ports Analysis and Labelling the Data

TABLE VI. PORT NUMBERS EXTRACTED FROM THE MKS DATASET

P2P Applications	Port Numbers
Skype	57290, 56091, 41900, 55303, 61976, 45220, 59774, 16130, 10131, 34625, 25406, 8999, 35133, ...
Funshion	64018, 1153, 48413, 52347, 38859, 56481, 13257, 29560, 48403, 4501, 61651, 54289, 44403, 29471, ...
Miro	50542-50544, 53778-53785, 61969, 61970, 50545, 53791, 51549, 51556, 50527, 37385, 51579, ...
BitTorrent	40283, 23791, 57605, 34923, 51084, 62383, 5530055308, 34319, 37192, 41011, 45177, 6771, 41843 ....
Tubi	50366, 50367, 50932, 50582, 50587, 50933-50939, 50596, ...
PPTV	50299, 50300-50310, 50073, 50072, 5041, ...
YuppTV	56213-56224, 55795, 56000, 55711, ...
AajTak	50975-50978, 64116, 64117, 50979, 50980-50984, 55548, 55549, 50809, ...
YouTube	54980, 54979, 50762, 50763-50767, 53026, 5076850771, 61049, 61082, ...
Vuze	13398, 50614, 57208, 57211, 57212, 57214, 57215, 57218, 57263, 57369, 57126, 57232, 59794, ….
BBC	52310, 52311-52315, 61196, 33419, 63738, 18340, 39701, 56727, 49183, 50270, 19702, ...
Hotstar	50489, 50490-50495, 50769, 61046, 61079, 61090, 63803, 63802, 50496, 50497, ...
Tribler	1130, 35140, 53736, 35190, 35175, 51122, 9206, 35120, 35130, 51044, 35080, 2105, 24934, 24935, ...
Gnutella	63432, 59650, 6602, 6791, 50088, 9216, 47655, 15398, 39961, 6312, 11553, 10381, 17983, 9812, ...
iQIYI	50486-50488, 50568, 50481, 50569, 50528-50530, 50570, 50571, 50475, 50505, 50428, 50533, ...

4. PERFORMANCE EVALUATION AND RESULTS

A. Experimental setup

As discussed in section 3, Wireshark is open-source software that effectively records the traces. The data collection is carried out using Desktop PCs with the processor @ 3.20 GHz in a Windows environment. The data are captured and saved as PcapNG files. The datasets are converted into CSV files for extraction of features. Features are extracted with GPU (4 Cores), with processor@3.80GHz, and 64GB Memory in Python environment. The extracted feature details are explained in TABLE IV. Chi-Square, ANOVA, and PCA are used to get a better feature subset. The ML techniques import the selected feature subset for identifying P2P traffic

from the given dataset. An exhaustive analysis is done considering different sets of features and with and without feature selection techniques. The detailed discussion on various methods used is given section 4C. The metrices used are given below.

B. Metrics for Performance Analysis

Internet traffic classification techniques require standard metrics to evaluate the desired goals by comparing the ground truth information. The following evaluation metrics are used to validate the proposed approaches:

- True Positive (T^+): the traces belong to P2P traffic and are classified correctly.
- True Negative (T^-): the traces do not belong to P2P traffic and are classified correctly.
- False Positive (F^+): the traces belong to a P2P traffic and are classified incorrectly.
- False Negative (F^-): the traces do not belong to a P2P traffic and are classified incorrectly.

The minimum value of F^+ and F^- indicates a good classifier. The metrics cited in [53] are used frequently for evaluation of the performance of the classifier with the help of T^+, T^-, F^+ and F^- is expressed as follows:

- Accuracy: This metric is used for evaluation of classification models. It is calculated by dividing the number of samples correctly classified positives and
negatives by total number of samples.

$$Accuracy = \frac{T^+ + T^-}{T^+ + T^- + F^+ + F^-} \quad (18)$$

Apart from accuracy, precision, and recall [54] are also used to assess the model, especially for the imbalanced classes. The details of these statistical measures are as follows:

- Recall: It is estimated as the ratio of correctly classified positives upon total positive count. Recall also called Sensitivity represents the percentage of overall positive cases present in the dataset.

$$Recall = \frac{T^+}{T^+ + F^-} \quad (19)$$

- Precision: It is the false positive rate or false alarm rate of a classifier which is estimated by the ratio of incorrectly classified negatives by the total negatives.

$$Precision = \frac{T^+}{T^+ + F^+} \quad (20)$$

- F1-score: This is Precision and Recall's harmonic mean.

$$F1 - score = 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} \quad (21)$$

C. Results and Discussion

As discussed above, the simulation was done in different ways by taking the different subset of input features. The simulation also done excluding the source and destination port as an input feature as well. The selected features using the feature selection techniques Chi-Square and ANOVA are mentioned in TABLE VII. The results are presented and discussed in subsections *C1* and *C2*. The comparative analysis is presented in subsection *C3*.

C1 Results considering the basic and flow-based feature subsets

TABLE VIII and IX listed the precision, recall, f1-score, and accuracy rate obtained after extensive simulation using the datasets of SAMPARK and UNIBS, respectively. The performance of the Chi-Square, ANOVA, and PCA techniques combined with different ML techniques are also analyzed and presented. The simulation results show that the performance of proposed hybrid model on SAMPARK dataset. The results show the significant improvements compared to the UNIBS dataset. The results show that the contribution of the Chi-Square method with most of the classifier outperforms other combinations. The accuracy rate achieved with the different combinations varies from 89% to 99%. From the results, it can be easily perceived that Chi-Square with Random Forest outperforms other approaches. By considering three feature subsets, the maximum accuracy achieved by this combination is 99.46%. The accuracy varies with the number of features. The selected features are indicated by the numeric number taken from the TABLE IV.

TABLE VII. FEATURE SELECTED USING FEATURE SELECTION TECHNIQUES

FS Method	Selected Features Name	
	For SAMPARK Dataset	**For UNIBS Dataset**
Chi-Square	Source Port (4), Destination Port (5), Flow Size (7), Flow Duration (9), 1st packet size in flow (8), Flow Count (6).	Flow Size (7), Flow Duration (9), Source Port (4), Destination Port (5), Flow Count (6), 1st packet size in flow (8).
ANOVA	Source Port (4), Destination Port (5), Flow Size (7), 1st packet size in flow (8), Flow Duration (9), Flow IAT (10).	Source Port (4), Flow Duration (9), Flow IAT (10), Destination Port (5), Flow Count (6), 1st packet size in flow (8).

The simulation was also done on UNIBS dataset considering same number of features. The selected features and results obtained are presented in TABLE IX. Here also Chi-Square, ANOVA and PCA is combined with the classifiers considered. It can be seen from the table that Chi-square or ANOVA with DT or RF performed better as compare to other models. It can also be inferred from TABLE IX that the maximum accuracy achieved is 96.06% for combination of ANOVA and DT when six features are considered. In case of three features the accuracy achieved is 94.52% for Chi-Square and RF model. Apart from the observations specified in the TABLE VIII and IX, it was tested further for more than six features on both the dataset SAMPARK and UNIBS, but the accuracy achieved is low and is not reported in the paper. For ease of understanding, Figure. 4, 5 and 6 graphically represent the comparison of classifiers in terms of accuracy of P2P traffic identification with different feature subset.

C2 Results considering the flow-based feature subsets with feature selection techniques

We also analysed the performance of the models by excluding the source and destination port from input features using the considerd feature selection techniques. The results are tabulated in the TABLE X and XI for the SAMPARK and UNIBS datset respectively. The ANOVA with RF classifier comparably performing better and achieved the high accuracy of 95.44% with four features on SAMPARK dataset. The selected flow-based features are Flow Size (7), 1st packet size in flow (8), Flow Duration (9), Flow IAT (10). Considering the five features (6-10), the Anova-RF and χ2-RF combination also achive the same accuracy. For UNIBS, as stated in TABLE XI, the ANOVA-RF combination is outperformed others with an accuracy of 90.89% considering the five features (6-10). The accuracy is good, but it is underperformed when compare the results obtained by simulation of model on SAMPARK dataset.

C3 Comparison with other existing approaches

The comparative analysis states that the proposed approach achieve better performance with an accuracy rate of 99.46% using Chi-Square as feature selection and RF as a classification technique on the SAMPARK dataset.

TABLE VIII. PERFORMANCE EVALUTION CONSIDERING SAMPARK DATASET

ML Techniques	**Precision**			**Recall**			**F1-Score**			**Accuracy**		
	$\chi2$	**ANOVA**	**PCA**	**$\chi2$**	**ANOVA**	**PCA**	**$\chi2$**	**ANOVA**	**PCA**	**$\chi2$**	**ANOVA**	**PCA**
Feature Selected: 02; Features are: (4, 5) for $\chi2$; (4,5) for ANOVA.												
Decision Tree	99	99	95	99	99	95	99	99	95	**99.40**	**99.40**	95.39
Random Forest	99	99	97	99	99	97	99	99	97	**99.40**	**99.40**	96.58
Naïve Bayes	90	90	88	89	89	82	86	86	84	89.09	89.09	81.71
KNN	99	99	97	99	99	97	99	99	97	99.12	99.13	96.74
SVM	98	98	95	98	98	95	98	98	95	98.48	98.48	94.90
Feature Selected: 03; Features are: (4, 5, 7) for $\chi2$; (4, 5, 9) for ANOVA.												
Decision Tree	99	99	97	99	99	97	99	99	97	99.24	99.13	96.53
Random Forest	99	99	97	99	99	97	99	99	97	**99.46**	99.08	97.34
Naïve Bayes	90	90	96	89	89	96	86	86	96	89.09	89.09	96.42
KNN	99	99	98	99	99	98	99	99	98	99.08	98.59	97.56
SVM	98	98	97	98	98	97	98	98	97	98.26	98.21	96.91
Feature Selected: 04; Features are: (4, 5, 7, 9) for $\chi2$; (4, 5, 8, 9) for ANOVA.												
Decision Tree	99	99	97	99	99	97	99	99	97	98.75	99.08	96.85
Random Forest	99	99	98	99	99	98	99	99	98	**99.13**	**99.13**	97.78
Naïve Bayes	90	90	97	89	89	97	86	87	97	89.09	89.42	97.02
KNN	99	99	97	99	99	97	99	99	97	98.53	98.59	97.34
SVM	98	98	97	98	98	97	98	98	97	97.88	97.88	97.12
Feature Selected: 05; Features are: (4, 5, 7, 9, 8) for $\chi2$; (4, 5, 8, 9, 10) for ANOVA.												
Decision Tree	99	99	97	99	99	97	99	99	97	98.81	98.53	97.18
Random Forest	99	99	98	99	99	98	99	99	98	**99.24**	99.02	98.10
Naïve Bayes	90	89	96	89	89	96	87	86	95	89.42	88.82	95.55
KNN	99	98	98	99	98	98	99	98	97	98.53	98.32	97.56
SVM	97	98	97	97	98	97	97	98	97	97.29	97.67	97.18
Feature Selected: 06; Features are: (4, 5, 7, 9, 8, 6) for $\chi2$; (4, 5, 7, 8, 9, 10) for ANOVA.												
Decision Tree	99	99	98	99	99	98	99	99	98	98.75	98.70	97.72
Random Forest	99	99	98	99	99	98	99	99	98	**99.13**	99.08	98.37
Naïve Bayes	90	88	95	89	89	95	87	86	94	89.37	88.82	94.90
KNN	99	98	98	99	98	98	99	98	98	98.53	98.26	97.88
SVM	97	97	97	97	97	97	97	97	97	97.23	97.45	97.23

TABLE IX. PERFORMANCE EVALUTION CONSIDERING UNIBS (UNIBS20091001) DATASET

ML Techniques	Precision			Recall			F1-Score			Accuracy		
	$\chi2$	ANOVA	PCA	$\chi2$	ANOVA	PCA	$\chi2$	ANOVA	PCA	$\chi2$	ANOVA	PCA
Feature Selected: 02; Features are: (7, 9) for $\chi2$; (4, 9) for ANOVA.												
Decision Tree	90	93	90	90	93	90	90	93	90	89.79	92.85	89.85
Random Forest	90	93	91	91	94	92	90	94	92	90.92	**93.66**	91.83
Naïve Bayes	85	85	83	88	88	87	85	85	84	88.03	88.17	87.49
KNN	88	91	91	89	92	92	88	91	92	89.30	91.69	91.77
SVM	88	88	89	89	89	89	85	86	86	88.86	89.11	89.29
Feature Selected: 03; Features are: (7, 9, 4) for $\chi2$; (4, 9, 10) for ANOVA.												
Decision Tree	94	93	92	94	93	92	94	93	92	93.68	92.71	91.59
Random Forest	94	93	93	95	94	93	94	93	93	**94.52**	93.64	93.23
Naïve Bayes	85	85	83	88	88	87	85	85	84	88.06	87.92	87.49
KNN	91	90	92	92	91	93	91	90	92	91.52	90.65	92.52
SVM	88	88	89	89	89	90	85	87	87	89.05	89.34	89.63
Feature Selected: 04; Features are: (7, 9, 4, 5) for $\chi2$; (4, 5, 9, 10) for ANOVA.												
Decision Tree	96	95	93	96	95	93	96	96	93	95.55	95.45	93.12
Random Forest	96	96	94	96	96	94	96	96	94	95.94	**95.79**	94.43
Naïve Bayes	85	85	83	88	88	87	85	85	84	88.07	87.97	87.48
KNN	96	95	93	96	94	93	96	95	93	95.60	94.50	93.29
SVM	90	89	89	90	90	90	88	89	87	90.33	9049	89.83
Feature Selected: 05; Features are: (7, 9, 4, 5, 6) for $\chi2$: (4, 5, 6, 8, 9, 10) for ANOVA.												
Decision Tree	96	95	93	96	95	93	96	95	93	95.64	95.44	93.22
Random Forest	96	96	94	96	96	95	96	96	94	95.84	**95.92**	94.55
Naïve Bayes	85	85	82	88	88	86	85	85	84	88.02	87.98	85.52
KNN	96	94	93	96	94	94	96	94	93	95.50	94.33	93.53
SVM	90	90	89	90	91	90	88	89	88	90.31	9078	9042
Feature Selected: 06; Features are: (7, 9, 4, 5, 6, 8) for $\chi2$: (4, 5, 6, 8, 9, 10) for ANOVA.												
Decision Tree	95	95	93	95	95	93	95	95	93	95.15	95.31	93.17
Random Forest	96	96	94	96	96	95	96	96	95	95.93	**96.06**	94.62
Naïve Bayes	85	85	82	88	88	86	85	85	84	88.04	87.91	85.56
KNN	96	94	93	96	94	94	96	94	93	95.54	94.49	93.57
SVM	90	90	90	91	91	91	88	89	89	90.64	90.79	90.56

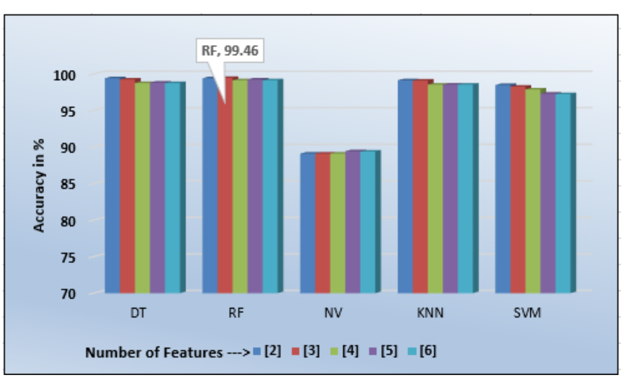

Figure 4. P2P identification accuracy of various ML Techniques applied on features selected using Chi-Square on SAMPARK dataset

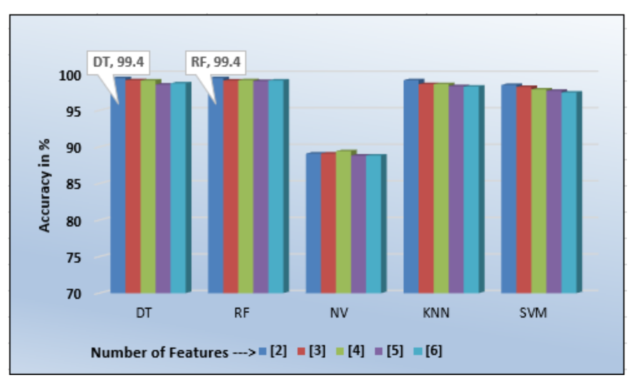

Figure 5. P2P identification accuracy of various ML Techniques applied on features selected using ANOVA on SAMPARK dataset

TABLE X. PERFORMANCE EVALUATION OF SAMPARK DATASET EXCLUDING THE SRC. AND DEST. PORT FROM INPUT LIST

ML Techniques	Precision			Recall			F1-Score			Accuracy		
	$\chi2$	ANOVA	PCA	$\chi2$	ANOVA	PCA	$\chi2$	ANOVA	PCA	$\chi2$	ANOVA	PCA
Feature Selected: 02; Features are: (6, 9) for $\chi2$; (8, 9) for ANOVA.												
Decision Tree	91	92	89	91	92	88	91	92	89	91.17	**92.19**	88,13
Random Forest	91	92	91	91	92	92	91	92	92	91.32	92.11	91.71
Naïve Bayes	74	89	91	86	88	90	79	83	88	85.93	87.52	90.01
KNN	74	81	74	86	85	86	79	81	79	85.93	85.46	85.93
SVM	86	91	92	88	92	92	86	91	92	87.95	91.79	92.44
Feature Selected: 03; Features are: (7, 8, 9) for $\chi2$; (8, 9, 10) for ANOVA.												
Decision Tree	95	91	91	95	91	90	95	91	91	94.54	90.77	90.17
Random Forest	95	93	94	95	93	94	95	93	94	**94.75**	92.91	94.18
Naïve Bayes	89	89	91	87	88	90	83	83	88	87.41	87.52	90.20
KNN	81	80	86	86	84	88	81	81	86	85.64	84.48	87.84
SVM	91	89	94	92	90	94	91	89	93	91.79	89.76	93.81
Feature Selected: 04; Features are: (6, 7, 8, 9) for $\chi2$; (7, 8, 9, 10) for ANOVA.												
Decision Tree	95	95	92	95	94	92	95	94	92	94.61	94.39	92.08
Random Forest	95	95	94	95	95	94	95	95	94	94.83	**95.44**	94.39
Naïve Bayes	89	89	91	87	87	90	83	83	88	87.37	87.34	90.16
KNN	81	80	86	86	85	87	82	82	86	85.75	84.59	87.48
SVM	92	89	93	92	90	93	92	88	93	92.08	89.69	93.09
Feature Selected: 05; Features are: (6-10) for $\chi2$: (6-10) for ANOVA.												
Decision Tree	95	95	92	94	94	92	94	94	92	94.39	94.39	92.00
Random Forest	95	95	94	95	95	94	95	95	94	**95.44**	**95.44**	94.28
Naïve Bayes	89	89	91	87	87	90	82	82	88	87.26	87.26	90.16
KNN	80	80	86	85	85	88	82	82	86	84.55	84.55	87.59
SVM	89	89	93	90	90	93	88	88	93	89.62	89.62	93.23

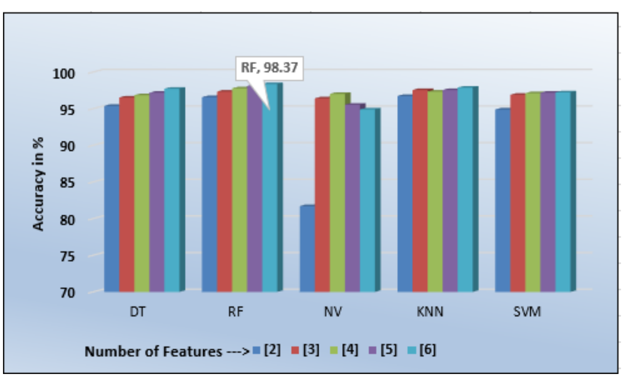

Figure 6. P2P identification accuracy of various ML Techniques applied on features selected using PCA on SAMPARK dataset

Yan et al. [25] achieved 93.9% flow accuracy on the UNIBS dataset using flow behavior-based technique. Jeffrey et al. [24] achieved 91.70% accuracy on the AUCK-IV sub dataset with limited port numbers. Saber et al. [28] have claimed the accuracy of 96% when it takes the shorter flow time-outs (15s) using PCA and SVM. Mohammadi et al. [29] have claimed a similar accuracy on his dataset but the approaches used are Genetic Algorithm and KNN classifier on a comparably small dataset. Wang et al. [40] and Cao et al. [23] has also achieved a good accuracy using feature selection and ML techniques. But Wang et al. used 10 ms sampling time of data packets whereas the dataset used by Cao et al. have mostly TCP (95%) data. Our SAMPARK dataset is large in size and consists of traces of traffic from varied protocols and services. Only 51% (appx.) traffic belongs to TCP.

The above comparative analysis reveals that the proposed approaches outperform the reported similar hybrid approaches. Among the proposed hybrid approaches, RF-$\chi2$ achieved the best accuracy on SAMPARK dataset. It is also noted that the feature selection techniques contributed to enhancing the performance of the proposed model. Random forest is mostly outperforming others and suited here because of large dataset.

5. CONCLUSION AND FUTURE WORK

Identifying traffic accurately has become one of the prerequisites for the network administrator to ensure adequate Quality-of-Service (QoS). The paper proposes

TABLE XI. PERFORMANCE EVALUATION OF UNIBS DATASET EXCLUDING THE SRC. AND DEST. PORT FROM INPUT LIST

ML Techniques	Precision			Recall			F1-Score			Accuracy		
	$\chi 2$	ANOVA	PCA	$\chi 2$	ANOVA	PCA	$\chi 2$	ANOVA	PCA	$\chi 2$	ANOVA	PCA
Feature Selected: 02; Features are: (6, 9) for $\chi 2$; (8, 9) for ANOVA.												
Decision Tree	89	84	85	89	85	86	89	85	85	89.23	84.67	85.51
Random Forest	90	86	87	90	88	89	90	86	87	**90.25**	88.10	88.78
Naïve Bayes	84	85	87	88	88	89	84	84	84	88.06	88.22	88.55
KNN	85	85	83	88	87	87	86	86	84	88.16	87.44	87.28
SVM	87	87	87	89	89	89	88	87	87	88.99	88.80	88.78
Feature Selected: 03; Features are: (7, 8, 9) for $\chi 2$; (8, 9, 10) for ANOVA.												
Decision Tree	89	86	87	89	87	88	89	87	87	89.19	86.71	87.56
Random Forest	90	88	88	90	89	90	90	88	89	**90.48**	89.38	89.72
Naïve Bayes	84	85	88	88	88	89	84	85	85	88.04	88.36	88.93
KNN	85	84	82	88	86	86	86	85	83	88.14	85.92	85.55
SVM	87	87	88	89	89	89	88	87	88	88.90	88.74	89.34
Feature Selected: 04; Features are: (6, 7, 8, 9) for $\chi 2$; (7, 8, 9, 10) for ANOVA.												
Decision Tree	90	89	88	90	89	88	90	89	88	90.03	89.14	87.84
Random Forest	90	90	89	91	91	90	90	90	89	**90.69**	90.63	90.11
Naïve Bayes	85	86	89	88	88	89	84	85	85	88.21	88.42	88.96
KNN	84	84	82	87	86	86	85	85	83	86.54	86.09	85.64
SVM	89	88	88	90	89	90	90	88	88	90.17	89.14	89.54
Feature Selected: 05; Features are: (6-10) for $\chi 2$: (6-10) for ANOVA.												
Decision Tree	90	90	88	90	90	88	90	90	88	90.19	90.19	87.82
Random Forest	90	90	89	91	91	90	91	91	89	90.69	**90.89**	90.15
Naïve Bayes	86	86	89	88	88	89	85	85	85	88.39	88.39	88.98
KNN	84	84	82	86	86	85	85	85	83	86.12	86.12	85.48
SVM	88	88	89	89	89	90	89	89	89	89.42	89.42	89.73

a hybrid methodology for P2P traffic identification. We have studied the effect of feature selection and ML methods for P2P traffic identification and proposed hybrid approaches by amalgamating port-based, feature selection, and machine learning techniques. The feature subsets are selected using Chi- Square, ANOVA, and PCA. The extensive simulation is carried out considering five ML algorithms and compared all the developed approaches with all possible combinations. The results have been analyzed and it is concluded that the Random Forest classifier with Chi-Square outperforms the other proposed approaches. The maximum accuracy achieved is 99.46% of accuracy and it is considerably better than the similar approaches reported in the literature.

It has been realized during the experimentation that P2P traffic identification is not sufficient while a fine-grained classification is emerging in the near future. It may also be important to establish a generic approach that can classify the new applications as well as existing P2P applications so that network traffic can be properly

handled. The research and SAMPARK dataset will provide a sound foundation for building and analyzing the solution for various P2P issues like selfish peer, flash crowd, overlay design, chunk scheduling, attack identification, etc.

REFERENCES

[1] Forouzan, A. Behrouz. Data communications and networking (sie). Tata McGraw-Hill Education, 2007.

[2] Thampi, Sabu M. "A review on P2P video streaming." arXiv preprint arXiv:1304.1235 (2013).

[3] Gomes, Joao V., Pedro RM Inácio, Manuela Pereira, Mário M. Freire, and Paulo P. Monteiro. "Detection and classification of peer-to-peer traffic: A survey." ACM Computing Surveys (CSUR) 45, no. 3 (2013): 1-40.

[4] Marfia, Gustavo, Giovanni Pau, Paolo Di Rico, and Mario Gerla. "P2P streaming systems: a survey and experiments." ST Journal of Research (2007): 1-4.

[5] Madhukar, Alok, and Carey Williamson. "A longitudinal study of P2P traffic classification." In 14th IEEE International Symposium on Modeling, Analysis, and Simulation, pp. 179-188. IEEE, 2006.

[6] Liu, Hui, Wenfeng Feng, Yongfeng Huang, and Xing Li. "A peer-to-peer traffic identification method using machine learning." In 2007 International Conference on Networking, Architecture, and Storage (NAS 2007), pp. 155-160. IEEE, 2007.

[7] Azzouna, Nadia Ben, and Fabrice Guillemin. "Analysis of ADSL traffic on an IP backbone link." In GLOBECOM'03. IEEE Global

[8] Telecommunications Conference (IEEE Cat. No. 03CH37489), vol. 7, pp. 3742-3746. IEEE, 2003.

[9] Pal, Kunwar, Mahesh Chandra Govil, and Mushtaq Ahmed. "Priority-based scheduling scheme for live video streaming in peer-to-peer network." Multimedia Tools and Applications 77, no. 18 (2018): 24427-24457.

[10] Pal, Kunwar, Mahesh Chandra Govil, and Mushtaq Ahmed. "Slack time–based scheduling scheme for live video streaming in P2P network." International Journal of Communication Systems 31, no. 2 (2018): e3440.

[11] Chen, Yishuai, Baoxian Zhang, Changjia Chen, and Dah Ming Chiu. "Performance modeling and evaluation of peer-to-peer live streaming systems under flash crowds." IEEE/ACM Transactions On Networking 22, no. 4 (2013): 1106-1120.

[12] Huang, Dan, Min Zhang, Yi Zheng, Changjia Chen, and Yan Huang. "Pre-allocation based flash crowd mitigation algorithm for large-scale content delivery system." Peer-to-Peer Networking and Applications 8, no. 3 (2015): 493-500.

[13] Biaou, Babatoundé O. Simon, Ayodeji O. Oluwatope, Helen O. Odukoya, Ajiboye Babalola, Oluwafolake E. Ojo, and Eric Herbert Sossou. "Ayo game approach to mitigate free riding in peer-to-peer networks." Journal of King Saud University-Computer and Information Sciences (2020).

[14] Bhatia, Max, and Mritunjay Kumar Rai. "Identifying P2P traffic: A survey." Peer-to-Peer Networking and Applications 10, no. 5 (2017): 1182-1203.

[15] Moore, Andrew W., and Konstantina Papagiannaki. "Toward the accurate identification of network applications." In International Workshop on Passive and Active Network Measurement, pp. 41-54. Springer, Berlin, Heidelberg, 2005.

[16] Saroiu, S., Gummadi, K. P., Dunn, R. J., Gribble, S. D., & Levy, H. M. (2002). An analysis of internet content delivery systems. ACM SIGOPS Operating Systems Review, 36(SI), 315-327.

[17] Baset, Salman A., and Henning Schulzrinne. "An analysis of the skype peer-to-peer internet telephony protocol." arXiv preprint cs/0412017 (2004).

[18] Reddy, Jagan Mohan, and Chittaranjan Hota. "Heuristic-based real-time p2p traffic identification." In 2015 International Conference on Emerging Information Technology and Engineering Solutions, pp. 38-43. IEEE, 2015.

[19] Roughan, Matthew, Subhabrata Sen, Oliver Spatscheck, and Nick Duffield. "Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification." In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, pp. 135-148. 2004.

[20] Karagiannis, Thomas, Andre Broido, Michalis Faloutsos, and K. C. Claffy. "Transport layer identification of P2P traffic." In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, pp. 121-134. 2004.

[21] Karagiannis, Thomas, Konstantina Papagiannaki, and Michalis Faloutsos. "BLINC: multilevel traffic classification in the dark." In Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications, pp. 229-240. 2005.

[22] Park, Byungchul, Youngjoon Won, JaeYoon Chung, Myung- sup Kim, and James Won- Ki Hong. "Fine- grained traffic classification based on functional separation." International Journal of Network Management 23, no. 5 (2013): 350-381.

[23] Cao, Jie, Da Wang, Zhaoyang Qu, Hongyu Sun, Bin Li, and Chin-Ling Chen. "An Improved Network Traffic Classification Model Based on a Support Vector Machine." Symmetry 12, no. 2 (2020): 301.

[24] Erman, Jeffrey, Anirban Mahanti, and Martin Arlitt. "Qrp05-4: Internet traffic identification using machine learning." In IEEE Globecom 2006, pp. 1-6. IEEE, 2006.

[25] Yan, Jinghua, Zhigang Wu, Hao Luo, and Shuzhuang Zhang. "P2P traffic identification based on host and flow behaviour characteristics." Cybernetics and Information Technologies 13, no. 3 (2013): 64-76.

[26] Raahemi, Bijan, Weicai Zhong, and Jing Liu. "Peer-to-peer traffic identification by mining IP layer data streams using concept-adapting very fast decision tree." In 2008 20th IEEE International Conference on Tools with Artificial Intelligence, vol. 1, pp. 525-532. IEEE, 2008.

[27] Draper-Gil, Gerard, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. "Characterization of encrypted and vpn traffic using time-related." In Proceedings of the 2nd international conference on information systems security and privacy (ICISSP), pp. 407-414. 2016.

[28] Saber, Abid, Belkacem Fergani, and Moncef Abbas. "Encrypted traffic classification: Combining over-and under-sampling through a PCA-SVM." In 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), pp. 1-5. IEEE, 2018.

[29] Mohammadi, Mehdi, Bijan Raahemi, Ahmad Akbari, Hossein Moeinzadeh, and Babak Nasersharif. "Genetic-based minimum classification error mapping for accurate identifying Peer-to-Peer applications in the internet traffic." Expert Systems with applications 38, no. 6 (2011): 6417-6423.

[30] Turkett Jr, William H., Andrew V. Karode, and Errin W. Fulp. "In-the-Dark Network Traffic Classification Using Support Vector Machines." In AAAI, pp. 1745-1750. 2008.

[31] Freire, Emanuel P., Artur Ziviani, and Ronaldo M. Salles. "Detecting skype flows in web traffic." In NOMS 2008-2008 IEEE Network Operations and Management Symposium, pp. 89-96. IEEE, 2008.

[32] Freire, Emanuel P., Artur Ziviani, and Ronaldo M. Salles. "Detecting VoIP calls hidden in web traffic." IEEE Transactions on Network and Service Management 5, no. 4 (2008): 204-214.

[33] Yang, Kai, Binqiang Wang, and Zhen Zhang. "A method of identifying P2P live streaming based on union features." In 2013 IEEE 4th International Conference on Software Engineering and Service Science, pp. 426-429. IEEE, 2013.

[34] He, Jie, Yue-xiang Yang, Yong Qiao, and Wen-ping Deng. "Fine-grained P2P traffic classification by simply counting flows." Frontiers of Information Technology \& Electronic Engineering 16, no. 5 (2015): 391-403.

[35] Perenyi, Marcell, Trang Dinh Dang, Andras Gefferth, and Sándor Molnar. "Identification and analysis of peer-to-peer traffic." Journal of Communications 1, no. 7 (2006): 36-46.

[36] Bashir, Ahmed, Changcheng Huang, Biswajit Nandy, and Nabil Seddigh. "Classifying P2P activity in Netflow records: A case study on BitTorrent." In 2013 IEEE International Conference on Communications (ICC), pp. 3018-3023. IEEE, 2013.

[37] Oudah, Hussein, Bogdan Ghita, and Taimur Bakhshi. "A Novel Feature Set for Application Identification." (2018).

[38] Junior, Gabriel Paulino Siqueira, Jose Everardo Bessa Maia, Raimir Holanda, and Jose Neuman de Sousa. "P2P traffic identification using cluster analysis." In 2007 First international global information infrastructure symposium, pp. 128-133. IEEE, 2007.

[39] Bhattacharya, Sweta, Rajesh Kaluri, Saurabh Singh, Mamoun Alazab, and Usman Tariq. "A Novel PCA-Firefly based XGBoost classification model for Intrusion Detection in Networks using GPU." Electronics 9, no. 2 (2020): 219.

[40] Zhongsheng, Wang, Wang Jianguo, Yang Sen, and Gao Jiaqiong. "Traffic identification and traffic analysis based on support vector machine." Concurrency and Computation: Practice and Experience 32, no. 2 (2020): e5292.

[41] Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set." Information Security Journal: A Global Perspective 25, no. 1-3 (2016): 18-31.

[42] Mahoney, Matthew V., and Philip K. Chan. "An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection." In International Workshop on Recent Advances in Intrusion Detection, pp. 220-237. Springer, Berlin, Heidelberg, 2003.

[43] Dusi, Maurizio, Francesco Gringoli, and Luca Salgarelli. "Quantifying the accuracy of the ground truth associated with Internet traffic traces." Computer Networks 55, no. 5 (2011): 1158-1167.

[44] Gringoli, Francesco, Luca Salgarelli, Maurizio Dusi, Niccolo Cascarano, Fulvio Risso, and K. C. Claffy. "Gt: picking up the truth from the ground for internet traffic." ACM SIGCOMM Computer Communication Review 39, no. 5 (2009): 12-18.

[45] Dong, Shi, and Raj Jain. "Flow online identification method for the encrypted Skype." Journal of Network and Computer Applications 132 (2019): 75-85.

[46] Verleysen, Michel, and Damien François. "The curse of dimensionality in data mining and time series prediction." In International work-conference on artificial neural networks, pp. 758-770. Springer, Berlin, Heidelberg, 2005.

[47] Su, Chao-Ton, and Jyh-Hwa Hsu. "An extended chi2 algorithm for discretization of real value attributes." IEEE transactions on knowledge and data engineering 17, no. 3 (2005): 437-441.

[48] Abdalla, Bushra Mohammed Ali, Haitham A. Jamil, Mosab Hamdan, Joseph Stephen Bassi, Ismahani Ismail, and Muhammad Nadzir Marsono. "Multi-stage feature selection for on-line flow peer-to-peer traffic identification." In Asian Simulation Conference, pp. 509-523. Springer, Singapore, 2017.

[49] Liu, Huan, and Rudy Setiono. "Chi2: Feature selection and discretization of numeric attributes." In Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388-391. IEEE, 1995.

[50] Jolliffe, Ian T. "Principal components in regression analysis." In Principal component analysis, pp. 129-155. Springer, New York, NY, 1986.

[51] Internet Assigned Numbers Authority (IANA),[Online: accessed on 10.06.2020] https://www.iana.org/assignments/service-names-port-numbers.

[52] Cotton, Michelle, Lars Eggert, Joe Touch, Magnus Westerlund, and Stuart Cheshire. "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name nd Transport Protocol Port Number Registry." RFC 6335 (2011): 1-33.

[53] Olson, David L., and Dursun Delen."Advanced data mining techniques." Springer Science and Business Media, 2008.

[54] Nguyen, Thuy TT, and Grenville Armitage. "A survey of techniques for internet traffic classification using machine learning." IEEE communications surveys and tutorials 10, no. 4 (2008): 56-76.

[55] CISCO VNI Global 2020 Forecast [Online: accessed July 26, 2020]https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast\-highlights/pdf/Global_2020_Forecast_ Highlights.pdf

[56] Fujita, Satoshi. "Flash Crowd Absorber for P2P Video Streaming." IEICE TRANSACTIONS on Information and Systems 102, no. 2 (2019): 261-268.

[57] Zhongsheng, Wang, Wang Jianguo, Yang Sen, and Gao Jiaqiong. "Traffic identification and traffic analysis based on support vector machine." Concurrency and Computation: Practice and Experience 32, no. 2 (2020): e5292.

[58] Abdalla, B. M. A., Mosab Hamdan, Entisar H. Khalifa, Abdallah Elhigazi, Ismahani Ismail, and M. N. Marsono. "Impact of Early Estimation of Statistical Flow Features in On-line P2P Classification." In 2020 IEEE Student Conference on Research and Development (SCOReD), pp. 294-299. IEEE, 2020.

[59] Bhatia, Max, Vikrant Sharma, Parminder Singh, and Mehedi Masud. "Multi-Level P2P Traffic Classification Using Heuristic and Statistical-Based Techniques: A Hybrid Approach." Symmetry 12, no. 12 (2020): 2117.

[60] Musa, Ahmad. "Analysis of UDP Traffic Norms through Packet Sniffing on Peer-to-Peer Networks." ATBU Journal of Science, Technology and Education 8, no. 2 (2020): 286-292.

Md. Sarfaraj Alam Ansari has received M. Tech degree in Information Security from NIT Durgapur, India. He is working as Assistant Professor and pursuing PhD at NIT Sikkim, India. He is having more than 06 years of industry experience and currently focused on academics and research. His research interest includes Computer Networks and communications, Information Security & Risk Management, Cloud computing, etc.

Dr. Kunwar Pal has received B. Tech in CSE from KNIT, Sultanpur, India in 2009, M. Tech (CSE) from PEC, Chandigarh, India in 2011 and his Doctorate from MNIT Jaipur, in 2018. Presently he is working as an assistant professor at NIT Sikkim. His research interests include communication, network security, real-time system etc.

Prof. M.C. Govil has received Doctorate degree from IIT. He is working as Professor at MNIT Jaipur, India. Presently he is on deputation at NIT Sikkim, India as Director. He is having more than 20 years of experience in education and research. He is member of various Societies. He has published many papers. His research interest includes communication, real-time system, software engineering etc.

Prajjval Govil is pursuing his B. Tech in CSE with specialization in Big Data Analytics from JK Lakshmipat University, Jaipur, Rajasthan, India. His research interest includes Machine Learning, Big Data Analytics and Speech Processing. He has completed more than 10 undergraduate projects on the topic mentioned above.

Adarsh Srivastava is an undergraduate student at NIT Sikkim, India. Presently, he is a sophomore and having a keen interest in Computer networks, Data Science. His further interests lie in solving complex problems using machine learning and deep learning.