Data Mining and Machine Learning Methods for Sustainable Smart Cities Traffic Classification: A Survey

Abstract

This survey paper describes the significant literature survey of Sustainable Smart Cities (SSC), Machine Learning (ML), Data Mining (DM), datasets, feature extraction and selection for network traffic classification. Considering relevance and most cited methods and datasets of features were identified, read and summarized. As data and data features are essential in Internet traffic classification using machine learning techniques, some well-known and most used datasets with details statistical features are described. Different classification techniques for SSC network traffic classification are presented with more information. The complexity of data set, features extraction and machine learning methods are addressed. In the end, challenges and recommendations for SSC network traffic classification with the dataset of features are presented.

Keywords: Sustainable Smart Cities, Security, Traffic, Classification, Data Maning, Machine Learning, A Survey,

1. Introduction

In this paper, we present the detail results of a literature review of sustainable smart cities, machine learning, data mining, datasets and its attributes for network traffic classification. The features extraction, selection and machine learning techniques as well as its applications are described to sustainable smart cities traffic classification. The different most widely used methods for sustainable smart cities (SSC) traffic classification are discussed with details. Furthermore, the complexity of the dataset, feature extraction, selection are

discussed [1]. The paper provides a set of comparison and recommendations for the best methods to SSC network traffic classification.

Internet traffic classification is a technique to identify and classify unknown network classes, or in summation, Internet traffic classification is the method to analyze and differentiate various class of applications flowing in a network. As T. Auld et al. in[2] discussed, that the method which helps the Network Operator (NO) in the identification of different types computer application that exist in a netowk [3]. Nowadays, SSC traffic identification becomes a scorching area in Sustainable Cities and Society (SCS), network management and network security, etc. Through this technique, various actions can be performed such as Network Operators (NOs) or Internet Service Providers (ISPs) can be able to handle the whole performance of a SSC network, monitoring, discovery, control, and optimization can be conducted on the classified traffic. Real-time network traffic classification has the capability to deal with the problem of sever network management for Internet Service Providers (ISPs) and their equipment vendors. It is essential for the SSC network operator to keep an eye on network [4], what is flowing on over their networks; thus they can manage and control the system in support of their goals. The end goal is to improve SSC network performance and identify intrusion or malicious of sustainable smart cities traffic. In our previous study, we studied sustainable smart cities and find out effective ML technique for Smart Cities (SM) traffic classification using ML technique [5] based on bijective soft set [6] .Furthermore, traffic classification is also the central part of the automated intrusion detection systems used to detect denial service attacks [4] reallocation of network resources or identify customer resource in a network and then reallocation. In the last decade, IP traffic classification technique is proposed, traditionally the methods have been based on direct inspection of each packet's contents of flows at a specific point on the network. Usually, simple Internet identification technique take place by using TCP or UDP port numbers which are visible in the TCP and UDP headers. But, mostly numerous applications are using random port numbers to prevent being detected[7]. Similarly, more complicated traffic classification method in-

fers application types mean well-known protocol behavior or application level signature by examining specifically available documentations within the TCP or UDP payload[8, 9]. Nevertheless, the effectiveness of the deep packet inspection is less impressive [10].

However, the research community, tries their best by investigating new SSC traffic classification schemes to classify application by inferring in application-level usage patterns without using deep packet payload inspections. Many new classification techniques to classify network traffic by using statistical features of the traffic flows such as mostly used statistical features packet lengths and inter-packet arrival times [11, 12]. Similarly, several researches endeavor hard to look application closely at the application of Machine Learning (ML) techniques, which is the subset of Artificial Intelligence (AI) discipline to IP traffic identification. However, the ML technique includes a number of some steps. Initially, traffics are capture and then feature (attributes) are extracted from the specific flow packets such as statistical feature maximum or minimum packet length, packet size, flow duration or inter-packet arrival times of each flow packets[13], etc. Then the machine learning classifiers are training with testing and training data set to classify unknown classes.

However, for SSC traffic identification, firstly traffic are traced, and then features are extracted after that detection method is applied to find out the normal and intrusion traffic. This survey paper, not only focused on machine learning and data mining techniques for sustainable smart cities (SSC) traffic classification as well as it also focused on Data sets and most crucial thing features which is very significant for for this technique to train the desire ML model for traffic classification methods and their descriptions. However, the methods that are described in this paper have been published after several reviews. Our paper includes the previously published papers reviews that are meet our article focused on methods and techniques.

This paper is handy for those who want to start their research study in the area of SSC Internet traffic classification using machine learning methods with an effective dataset and their features. This paper not only includes ma-

chine learning for network traffic classification. A dataset is also very utmost for network traffic classification, so the datasets and its useful features are also discussed and described in this paper. As a great, thorough description of the ML/DM technique and dataset of features for anomaly and intrusion network traffic classification are placed.

Radhika et al.[14] describe the machine learning based network traffic classification system. Teodoro et al. [15] review most popular used sustainable smart cities traffic identification technique. Their paper includes statistical[16], knowledge and ML techniques. But their study cannot give detail state-of-the-art ML/DM techniques and datasets. Nguyen et al. [17] describe a ML technique for Internet traffic identification. Their survey consists of the paper published from 2004 to 2007. Similarly, Anna L et al. [18] describe machine learning and data mining methods for cybersecurity intrusion detection. Their survey includes only machine learning, data mining and dataset. However, they mostly focused on machine learning and data mining method for cybersecurity intrusion detection. Moreover, their survey paper only covers published paper about cybersecurity intrusion detection, whereas our paper cover machine learning, dataset and its features extraction and effective selection method for Internet traffic identification. Unlike Anna L et al. [18], introduce the methods and technique that can be used for Internet traffic classification using machine learning classifiers not only cyber intrusion detection. Callado et al.[19], describe the different methods for Internet traffic classification such as Signature-based packet payload analysis. However, their survey only limited to network traffic classification technique. While our survey paper covers all the different techniques used for Internet traffic classification as well as different methods and different dataset features extraction and selection techniques. Bhatia Max et al.[20] present identifying Peer-to-Peer (P2P) traffic survey paper. Their article includes several strategies in the identification of Peer-to-Peer (P2P) traffic and also present a detail theoretical analysis for the measurement and monitoring of a network. They also describe some network traffic classification technique, but their survey also limited to Peer-to-peer (P2P) and encrypted traffic. Moreover,

they do not present details on ML/DM and dataset used for Internet traffic identification.

In this survey paper primarily we focus sustainable smart cities traffic classification then dataset, which is very important for classification and then feature extraction and selection, while after detail description of machine learning different Internet traffic classification are described.

The rest of the paper is organized as follows: In Section II we provides an overview of network anomaly traffic classification, Sect. III gives necessary steps for SSC traffic classification using ML and DM methods. Sec IV provides dataset used for anomaly and normal internet traffic classification. Similarly, Sec V introduces the in-depth feature extraction and selection. While In Sect VI detail machine learning and data mining technique for SSC internet traffic classification are shown. Sect VII presents the details methods for early stage internet traffic classification with the framework. Sect VIII presents observation and recommendations, while in Sec IX Conclusion.

2. Key Techniques of Sustainable Smart Cities traffic classification

The topic of sustainable smart cities traffic identification or network traffic identification achieved much importance recently in the scientific research contributions due to several aspects related with it, like intrusion detection, network traffic management, network security and quality of services (QoS), etc. An accurate Internet traffic identification in a massive network environment is utmost necessary for Internet Protocol (IP) network management, network security, monitoring and quality of services (QoS) of, etc. However, due to network traffic classification, a network operator or Internet service providers (ISPs) can handle and control the performance such as blocking unwanted traffic flow in a network and managing resources, etc. Moreover, due to these techniques, ISP or Network Operator (NO) can easily find out the growth of network applications and then can manage some resources for the desired applications which are growing day to day.

As Hurley J et al. [21] discuss the growth of the applications traffic day today and showed that its consume large size of bandwidth such as P2P applications traffic consume considerable size bandwidth in the network. The main question arises here, why these applications consume large size of bandwidth such as P2P application traffic. However, due to bidirectional flows traffic at the same time these applications traffic consume large size of bandwidth. Thus it is clear that if the flows of traffic bidirectional it will consume large size bandwidth. However, due to these reason P2P applications consume large size bandwidth. Moreover, not only Peer-to-Peer application takes place in consuming large size of bandwidth in available network bandwidth as well as SMTP, FTP and HTTP, etc. network applications also consume large size of bandwidth. It means that it is vital for Internet service provider and network operator to keep an eye on these problems as well as to implement high Quality of Services (QoS) policies [22] for each application respectively to maintain the performance of network effectively. Similarly, ISP is also facing numerous problems such as satisfying customer with effective quality broadband experience, expensive backbone links, and upstream bandwidth, etc. The authors in [23, 24] shows that P2P applications consume 60% bandwidth of the total available bandwidth which holds mostly portion of available Internet bandwidth. H. Schulze[25] conducted a study on worldwide network traffic measurement and showed that mostly file sharing application such as P2P application produce a number of traffic in comparison with other Internet application network traffic. In their study, they classify Web browsing, P2P file sharing, Internet telephony, online game, media streaming and instant messaging, etc. They also showed that web traffic got comeback popularity due to social networking, file sharing and day by day growing media Web pages. However, this is extreme crucial for ISPs to maintain the quality of services (QoS) of those applications which generate more and more traffic. Not only QoS while it is also difficult for ISPs to implement network security and intrusion detection for each flow traffic in a massive network. However, traditional traffic identification technique covers the problem of classification and identification of several different applications traffic flows to manage the overall

performance of Internet traffic flows accurately and conduct effective security methods. Nevertheless, it is vital to manage these types of traffic effectively to ensure the network performance. Traditionally IP traffic classification includes on the inspection of packets of TCP or UDP port numbers means Port-base classification or Payload-based traffic classification. The necessary demonstrations are given below of some traditional traffic classification techniques are shown in figure 1.

2.1. Port based traffic classification technique

TCP and UDP generate the several flow connection communication using port numbers between common IP endpoints. In [17] studied that many application use 'known' port number for their local host communications. Similarly, an identification classifier is sitting in the center of network for the purpose to view at the TCP SYN packets and to know the TCP connection of new client-server on the service side. However, the TCP SYN mean the initially steps in TCP'S for session establishment using three-way handshake [26]. Then the application in inferring by viewing the port number of packet in the registered list of Internet Assigned Numbers Authority (IANA)'s [27]. In simple words, the network application first registers their port number with IANA list and then using the register port number the corresponding traffic were identified. For instance, the E-mail application to send email Simple Mail Transfer Protocol (SMTP) use 25 well-known port number while for receiving email Post Office Protocol (POP3) used 110 port numbers. Similarly, web applications use 80 well-known port numbers.

Though, Port-based technique is very useful and great for classification/ identification of network application in a massive traffic network. However, this technique has some limitations such as mostly application used a dynamic port number for their communications like P2P application and many applications don't use port number instead of dynamic port number same as P2P, Napster, and Kazaa [7, 28]. While dynamic port number means unregister port number

with IANA. In other words, to avoid being detected and access control of the operating system an application may use a different port number in contrast of a widely used port number. Similarly, Real Video streamer dynamic [29] port number of server port for the data transfer and then the initial TCP connection the server port is negotiated and then established RealVideo port. Moore and Papagiannaki studied that this approach cannot get over 70% accuracy results for port-based traffic identification and does not provide effective classification accuracy [30, 31]. Madhukar and Similarly Sen et al. [9] studied that port-based traffic identification only classifies 30% of Kazaa P2P application.

However, port-based technique failed due to the use of the dynamic port number.

2.2. Payload based traffic classification technique

To address the problem of port-based technique, a new technique named payload based approached is introduced. This method is also known as Deep Packet Inspection (DPI) technique. In this technique, the content of the packets are analyzed the characteristics signatures of the network applications traffic. This technique is the second method after and alternative approach of port-based technique. However, this method especially produced for P2P type's applications traffic classification which used dynamic port numbers instead of a port number. Karagiannis et al. in [7] and Sen et al. [9] studied that analyzing the signatures of the traffic at the application level mean payload based classification of P2P traffic can reduce 5% false positive and false negative for P2P traffic. Likewise, Moore and Papagiannaki in [30] combine use both port-based and payload based method for the identification of network applications. They start their classification procedure with the port number to analyze the flow first. If in classification no known port number discovered then the traffic is passed to another next step. While the second step includes on examine to see flow contains the signature or not. If no one technique better to identify, then the packet is analyzed to look if the packet includes a protocol. If these both tests fail then the flow of the first Kbyte of protocol signatures are studied.

However, their result shows that port information is able to classify Internet traffic about 69% of the total bytes. They also show that the information of the first each flow can lead to increases the accuracy of approximately 79% in classification. However, this technique is not very suitable and sufficient for Internet traffic classification. The big problem in this approach is that this method needs costly equipment for pattern checking in a payload. While the another big issue is that the method does not classify encrypted applications traffic. And the third one is that this method require ongoing updates of signature pattern for the fresh coming applications. Although this technique is the substitute to the port based technique, and is very complicated as compare to port based technique. These techniques violate the user and organization privacy policy.

2.3. Statistical traffic properties based classification

To overcome the port-based and payload based problems a new statistical traffic properties based method classification method proposed this approach based on statistical characteristics of traffic to identify or classify the applications in a network. While statistical traffic characteristics are the network layer properties such as flow duration, packet length, packet inter-arrival time and flow idle, etc. which are very useful for the network traffic identification to distinguished various types of application in a network. V.Paxs in [32] found that the relationship between the traffic and studied its statistical characteristics or properties [33] and also analyzed as well as construct an empirical model for connection characteristics like bytes, durations and arrival periodicity for a specific TCP application. In their study, they showed that their model could describe a details description as well as empirical models. However, their study limited to TELNET, FTP, SMTP, and NNTP applications connections. Similarly, Tanga Lang et al. in [34] develop a systematic traffic model for the online computer game Half-Life. The primary goal of their research is to estimate the potential future effect of the online game on the IP network. For their study, they observed and characterized packet length, packet inter-arrival times and rates mean packets and bites per second. In [35] the authors developed a system-

atic traffic system for the online computer Quake3 game. They used the same method that in above used, but the traffic classification environment is different [36]. They applied for their study the same traffic statistical characteristics which they have used in the previous paper[37]. However, these classification methods are based on traffic flow statistical properties.

To deal with the problems of traditional traffic classification machine learning (ML) based traffic classification technique proposed, which is very effective with respect for the network traffic identification in a large network traffic environment. In this survey paper, we describe with very details the necessary steps in machine learning and data mining techniques. In 2015 Anna Buczak and Erhan Guven in [18] studied that there is plenty chaos in ML, DM and Knowledge Discovery in Database (KDD). They considered that KDD is the step by step process to extract the useful previous information from the input data. Similarly, Fayyad U et al. [38] studied that knowledge discovery in database is the term indicate the whole step by step method of discovering knowledge from the data. They also studied that, in KDD machine learning is the very important step and also for pattern extraction. They also discussed that in this method there are some important steps such as making data and then selection, data cleaning, initially knowledge and interpretation of appropriate knowledge, and the last one is result. But this process guarantee that extracted information is extracted from the data. Nevertheless, in [39] the author calls the whole process of knowledge discovery in database and data mining. ML and DM are two confused terms in the area of computer science. In 2015 [18] the authors state that these terms are very confusing and overlap with each other and often used the same method. Similarly, Authur Samuel studied and defined that machine learning as the study that make computers able to learn without any overtly programmed. However ML keep eye on prediction and identification based on known pattern which is being learn from the training data. The major dissimilarity between DM and ML is that machine learning needs problem formations from the domain main dependent variable to predict. While data mining attentions on the detection of previous undiscovered properties in the data [18]. Data

mining does not require a particular goal mean problem formation but keep attention on searching new and interesting information (knowledge). However, the term Data Mining (DM) was brought in the late 1980s in the first KDD conference and place in 1989s while the term ML has been used in since 1960s. While new researchers use the term DM instead of ML because of its popularity. Anyway, we only concentrate in this survey on ML classification. ML technique includes of two main steps: training and testing steps. But the overall traffic classification phases are shown in Figure 1. This is just like a short tutorial on network traffic classification using the ML technique. But particularly in ML sections, the class attributes mean features and class are identifies then the subset of attributes are identify for identification then the model learn by using training data and then by using the trained model the unknown data is identify or classify. The authors in [18] shows in their study that in case of misuse identification the misuse classes then learned by using suitable examples from the training data set. In the testing phase, the data are run by model and instances are classified whether its belong to the misuse class or not. However, in internet traffic classification firstly traffic sample are defined within the training set phase. While in testing phase the earn model is applied to the upcoming new data and similarly every instance is classified as either related to concern class or not. Nevertheless, the most important steps necessary for Internet traffic classification are described with details below as shown in Figure 1.

- Network Traffic Capturing: In network anomaly traffic classification technique. The most important and necessary step is tracing network traffic to make effective datasets for effective anomaly internet traffic classification. In this phase, the real Internet traffic is traced using different types of tracing applications such as Wireshark [40], tcpdump tool [41], Snort and Nmap [42], etc.

- Feature Extraction and Selection: This step includes features extraction and selection for the ML technique. Without feature extraction, it is impossible to perform network traffic identification by using ML method.

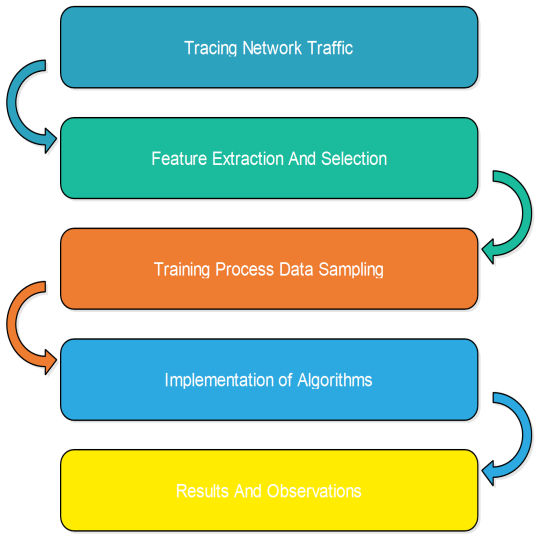

Figure 1: Basic Steps in Traffic Classification Technique

In this method, the feature are created from the trace traffic like packet size, packet length, packet duration, packet inter and arrival time, etc. Using the extracted features machine learning classifiers is trained. For features extraction from the traced traffic, there are many online publically available software applications such as Perl script can be employed to creat or extract features and another one NetMate tool [43] which extract almost 44 statistical features from the flow traffic.

- Training Process or Sampling: In this step, the data set is sampled using a supervised learning technique. The samples are initially labeled to identify unknown class's traffic. In other words, in this step data set are divide into training and testing data sets for the implementation of machine learning (ML) classifiers. Though all steps are very important but this step is essential, because without sampling or making training and the testing dataset it is impossible to classify or identify Internet traffic effectively.

- Implementation of ML Algorithms: In this step, the data set are of training, and testing sets are conducted using selected ML classifiers. for the

implementation of ML classifiers, there are much open access software application such as MatLab [44] and Weka classification simulation tools [45].

- Results and Observation: After the execution of ML classifiers, the application tool that is used for simulation provide a details results information such as accuracy details information, training and testing time and recall [46], etc.

However, our survey paper follows the given model as shown in Figure 2.

3. DATA SETS FOR TRAFFIC CLASSIFICATION

In Internet traffic classification, the data is fundamental. Because without data it is impossible to conduct Internet traffic classification technique as well as to understand the Internet traffic identification technique it is vital to understand the data that will use in network traffic classification technique using machine learning (ML) methods. It is also utmost important to know before starting the classification/identification technique using ML/DM technique different researcher authors utilized various types of data sets for their research study. However, in this section we describes in details the used datasets specifically in network traffic classification technique using machine learning classifiers. Basically, some well-known dataset are also describe in this section. In 2010 L.Peng et al. in [47] Studied that mostly researcher in the area of machine learning based internet traffic classification follow the traditional way such as collection of sample on key node of the network for their research study. They studied that traditionally collected samples don't carry enough application information for identification like ground truth, which is very important for ML algorithms in network traffic classification. However, they developed and designed distributed host-based traffic collecting platform (DHTCP) for the tracing accurate traffic on user hosts [48]. By using the proposed platform they build a dataset. Moreover, for accurate tracing traffic they conducted DHTCP: TSLSP, traffic sensor,

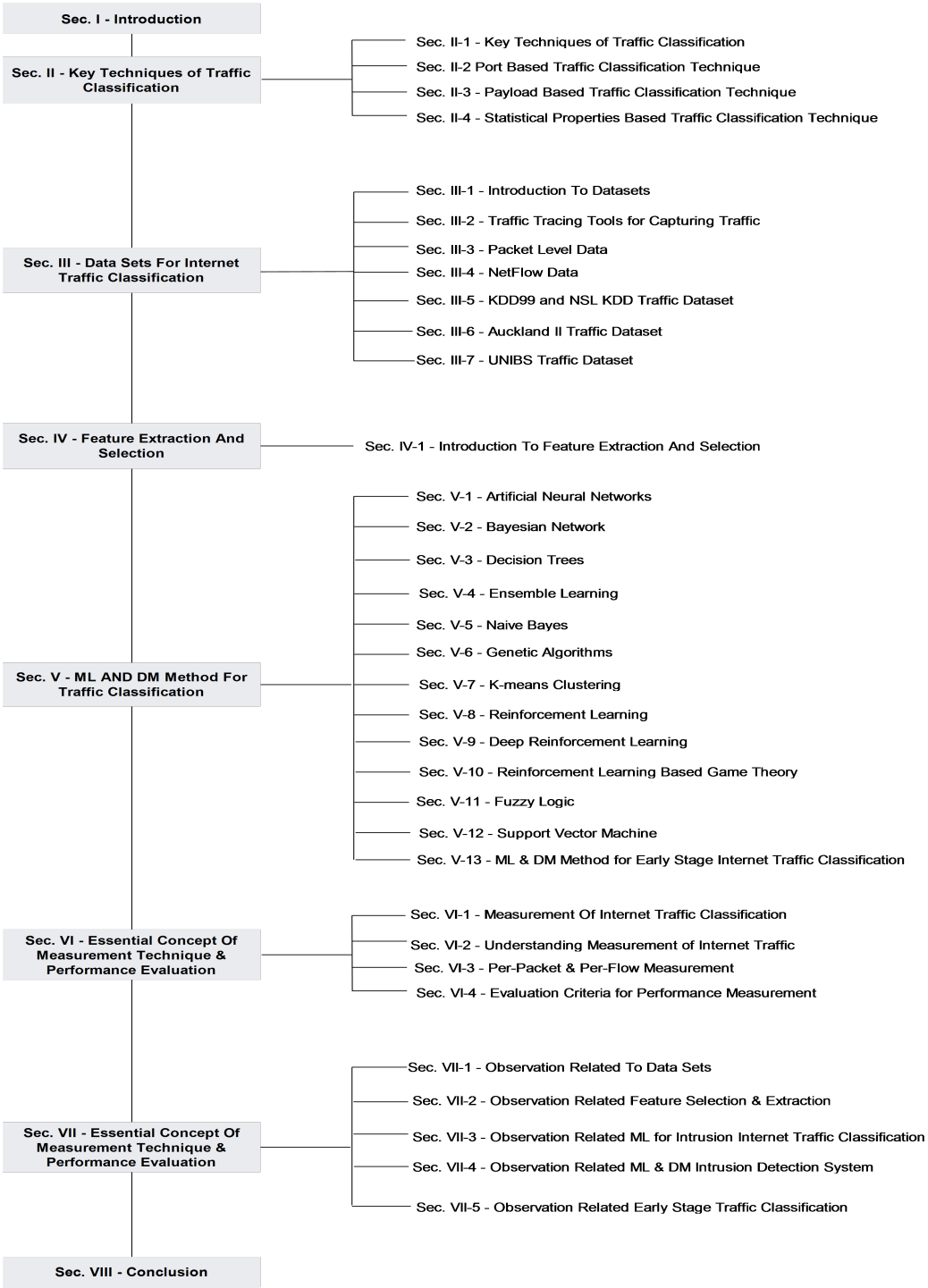

Figure 2: Proposed Survey Model

TS server and data analyzer. They apply deploy the proposed technique Harbin Institute of Technology and the University of Jinan respectively. However, they mention that it is very crucial to traced accurate traffic data for Internet traffic classification. Bellows are some datasets and their details information. In Section 3.1 initially, we will discuss the tools used for capturing traffic and then gradually goes to in details datasets information.

3.1. Traffic Tracing Tools for Capturing Traffic

The critical task in network traffic classification is capturing traffic accurately and carefully to protect user privacy and other sensitive information. In 2007 Mark Allman and Vern Paxson in [49] published that note on the care required when releasing measurement data. They propose a guide and help for how to treat data measuring. However, traffic can be traced by the help of tcpdump [50], WinDump [51], WinPcap [52] libraries, and Wireshark tool [40], but using these tool application take much ample space to store the captured data. However, to minimize the size of trace traffic data trace reduction method is used for this type of problem using packet filtering method. In more details to reduce the burden of trace traffic data by tracing only the limited amount of traffic, not full flow traffic. Another technique is storing only the TCP/IP protocol header information. However, mostly researcher community used online publically available dataset for a research study such as Auckland II data set publically available in [53] and another mainly used dataset is UNIBS data set which also publically free available on the internet in [54]. However, some datasets level are discussed below.

3.2. Packet Level Data

The authors in [18] shows that the applications that are used by users are generates the packets network traffic can be capture or trace by using different types of application such as to trace the packets that are received and transmitted at physical interface by using WinPCap and Libpcap and windows version. While the most widely used applications used in windows are tcpdum

[41], Wireshark [40], Snort and Nmap [42]. In more depth, an Ethernet frame is possess on Ethernet frame of the Ethernet header called (MAC) at physical layer about fifteen hundred bytes of payload.However, the payload consisted of an IP packet, which is consist of transport layer IP header and the IP payload. Callado in [18] state that Internet Protocol (IP) payload includes on data or in other word other protocol which is highly encapsulated and they give details that the packets can be traced by using pcap interface and that features are varies with protocol that carries packets.

3.3. NetFlow Data

NetFlow data can be collect through router or switch. In the computer network enter or exit of traffic flow can be trace or capture with help of router or switch. It's mean, switch or router can trace the traffic flow in a network. Cisco's NetFlow version 5 describes a network flow as a unidirectional sequence of packets that share seven packet features such source IP address, destination IP address, source port, destination port, IP protocol and IP type of services. However, the NetFlow architecture includes on three components such as a NetFlow Exporter, a NetFlow Collector, and Analysis Console. Nowadays, there are ten NetFlow versions. NetFlow Version 1 to 8 are same to same. However, the NetFlow data consist on the compressed and preprocessed version of the actual network packets, and the statistical features are then extracted which are based on some attributes like the number of packets, duration of window, etc.

There are many data sets for researchers which are publicly available on the Internet, but we select the most used some of them which are very important in the ares of network traffic identification using ML algorithms.

3.4. KDD99 and NSL KDD Traffic Traces Data set

KDD99 is the first benchmark dataset of Intrusion Detection System was developed by DARPA. In this data set, various types of anomalies attacks are categorized into four different groups like Probe, DOS, R2L, and U2R in [14].

This dataset includes on forty-one features, which consist of normal and attacks traffics. Nowadays, many researchers used the dataset for IoT anomaly traffic classification to effective classify anomaly and intrusion attacks in IoT traffic. In 2015 Anna Buczak and Erhan Guven in [18] studied that there is many chaos in ML, DM and KDD. They studied that KDD is the step by step process to extract the useful previous information from the input data. Similarly, Fayyadi U et al. [38] studied that KDD is the term indicates the complete process of discovering knowledge from the input data. They also studied that DM is the specific phase in the KDD process and models from data. However, the steps in the KDDD process are data creation, data selection and cleaning, knowledge extraction, and proper interpretation of knowledge, and finally results of DM that are true. Nevertheless, in [39] the author calls the whole process of KDD DM. Though the dataset is widely used, but dataset has some limitation which is pointed out in [55]. However, to avoid the limitation NSL KDD dataset was introduced which has the advantages in [56]. To decrease the classifiers bias the redundant records were deleted from the dataset. The duplicate was, and new train and test sets were created.

3.5. Auckland II Traffic Traces

Waikato Internet Traffic Storage project collect and document all the Internet traffic. In which some of traces dataset are freely available for researchers. Though there are a lot of datasets available for anomlay network traffic identification such as Auckland I, IX, IV, and VI, etc, but we choose only Auckland II, because Auckland II data set is very accurate for Internet traffic classification and use by many researchers for their research studies. Auckland II trace traffic is the collection of long GPS Traces at the University of Auckland using pair of DAG 2 which is available at [57]. This data set consist on 85 different traces files which are captured in November 1999 to July 2000. However, Auckland II data set were set focused on 24-hour runs. But due to breakdown of some hardware, the trace become shorter. But in these traces traffic the two trace traffic files traced at very accurate trace traffic for network traffic identification by using

ML algorithm in the best our knowledge. These two trace traffic files includes on header bytes for each frame. But in these dataset the payload of application is not presented as well as the complete IP addresses are not known by using Crypto-Pan AES encryption. Similarly, tracing Auckland II traffic, the research team used DAG3.2E card with 100 Mbps to trace the traffic at border router of University firewall. However, they did not trace the payload of application and they studied that DPI tools are not effective for the ground truths. Nevertheless, only port numbers is effective to trace the original application type. In 2015 Lizhi Peng et al. [58] used Auckland II data set for their research Internet traffic classification. In their research work, they use Auckland II and selected eight different types of application from the Auckland II trace traffic and then they conduct filter process on the traffic flows with mouse flows with non-zero packets. The details information of the trace traffics are shown in the given Table 1.

Table 1: AUCKLAND II TRAFFIC TRACES

Type	No Instances	Bytes
ftp	251	137241
ftp-data	463	5260804
ehttp	23721	139421961
imap	193	86455
Pop3	498	98699
smtp	2602	1230528
nntp	274	22108
ssh	237	149502
DNS	5488	511137
telnet	37	21171

3.6. UNIBS Traffic Traces

UNIBS is freely available traffic dataset for Internet traffic classification traced by Prof. F. Gringoli with his research team, which is freely available at [59]. Their research teams develop a very useful system application namely GT. The traffic were traces by using edge router at University of Brescia of the University campus during only the three days (September 30, October 1 and 2 2009) and collected the traffic by using Tcpdump [50] on the faulty router, which is connected to the network with a 100Mb/s uplink [60]. While the flows are 99% includes on TCP flows. However, mostly researcher related to Internet traffic classification conducted shown in the given Table 2.

Table 2: UNIBS TRAFFIC TRACES

bittorrent	251	6393487
Edonkey	379	241587
http	25,729	107342346
Imap	327	860226
Pop3	2,473	4292419
Skype	801	805453
msn	60	3753
Smtp	120	43566
Urd	650	132209
ssh	23	39456

4. FEATURE EXTRACTION AND SELECTION

In Internet traffic identification the biggest crucial phase is feature extraction and selection. Without feature extraction and selection it is impossible to classify and identify anomaly and unknown classes. Thus understanding feature

selection and extraction is utmost necessary for the field of network traffic identification. Considering Internet traffic classification to the best of our knowledge several studies are conducted to improve the performance of identification by through machine learning algorithms. However, the proposed models are very promising to get very accurate identification accuracy more than 80% by using several different network environment datasets. But to classify more effective and precisely mostly researcher faced the imbalance traffic problem in network traffic identification. However, the class imbalance traffic and concept drift is a huge problem in Internet traffic identification. To solve the problem of class imbalance and concept drift the researcher endower hard to propose effective approaches, in which feature selection and extraction methods got very importance for imbalance traffic identification. However, features selection is very crucial in network traffic identification. Lim et al. in [61] investigate the features and analyzed the effectiveness of its class distribution[62]. Similarly, Wasikowski and Chen in 2010 [63] studied and develop three different methods for internet traffic classification and then compared seven different types of metric by considering the class imbalance problem. There proposed methods used for feature selection. In their study they evident that correlation coefficient of signal noise and Feature Assessment by Sliding Thresholds (FAST) are important for accurate feature selection particularly for imbalance traffic classification identification. However, it is necessary to select and proposed effective feature selectin technique for effective network traffic identification. For this purpose More et al. [64] proposed valid attributes selection and attributes extraction techniques for network traffic classification which is used by numerous researchers for their study. They extracted 248 flow statistical features from the total flow traffic such as packet size, average packet size, maximum and minimum statistical features, etc. By using above given features, they achieve auspicious results for the classification of network traffic.

Recently, Hongli Zhang et al. [65] Presented two different algorithms for effective feature selection. Using the proposed algorithms, they select accurate feature with AUC metrics for Internet traffic identification. For the proposed

algorithm performance evaluation they used true positive rate (TPR) and false positive rate (FPR) and evident that their proposed method is able to get more than 90% accuracy for network traffic identification. Similarly, Lizhi Peng et al. [58] showed and verify the different features set effectiveness at early stage network traffic classification. classification. They showed in their study that early stage feature of packet carries enough classification information at early stage Internet traffic identification. In more depth Bernaille et at. [66] Study the problem of feature selection and evident packet size as a feature. They also used the extraction technique and extract few statistical attributes from it. While, they used K-means, GMM and HMM model for network traffic identification [67]. Similarly, Lim et al. in [61] derived few features from the packet size and then used the extracted set of features with connection level for network traffic identification. But, their still need more improvement in effective feature selection.

As we discuss, the most important and crucial task in network traffic classification is features extraction from the trace traffic. For effective traffic classification using machine learning techniques effective features are also very important. In 2012 Ding S et al. [68] discuss the general discussion of feature extraction. In which they discuss linear and nonlinear feature extraction, but the author and his friend discuss features extraction related to pattern recognition. Another complicated issue in Internet traffic identification is the selection of effective feature for internet traffic identification. In 2012 Zhang H et al. [65] address the problem of features selection and they proposed Weighted Symmetrical Uncertainty (WSU) and developed a feature selection algorithm WSU_AUC and then they propose SRSF algorithm that are able to select stable features from the result got by WSU_AUC. Similarly, in 2014 Chen Z et al. [69] demonstrate that Effective features for accurate network traffic classification with respective time and location are sensitive. For accurately feature selection in 2015 Bennasar M et al. [54] proposed two nonlinear methods and showed that their two methods are effective.

5. ML AND DM METHODS FOR TRAFFIC CLASSIFICATION

This section demonstrates different ML and DM methods for anomaly and intrusion traffic classification. However, we try our best to describe each technique with details. First, we will discuss literature related to intrusion detection and then goes deep in ML and DM technique. We also give references to works that are related to the method and some specific papers that are proposed methods to Internet traffic classification.

This section especially presents the methods used by researchers for anomaly

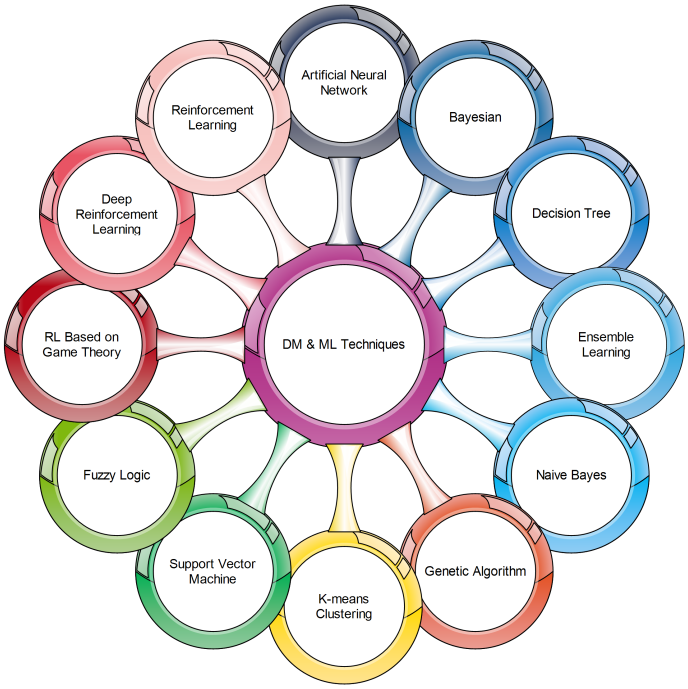

Figure 3: DM and ML Techniques

and intrusion detection systems in Internet traffic classification as shown figure 3. Diro at el. [42] studied the problem of attack detection and proposed a deep learning attack detection system by using different well-known dataset such as KDDCUP99, NSL-KDD and ISCX for attack detection in a computer network. They used accuracy, false alarm, and detection rate metric to find out

the proposed approach performance effectiveness. Recently Acharjya et al. 68 introduced a method to identify the exact activity based on Motion Projection Profile (MPP). They extracted features from a temporal difference image for the various levels of a person's posture. By analyzing the projection profile features fall are detected consist of each row motion pixel, column, left diagonal of the images difference stretches information to distinguish prompt posture of the person. For the experimental analysis, they used publically available fall detection dataset and extracted MPP features set. Moreover, they used various well-known machine learning algorithm like SVM with the polynomial kernel, SVM, KNN and Decision tree (J48) algorithms for the classification of fall activities. From the experimental analysis, they showed that SVMRBF kernel gives promising results for the fall detection which gives 89.55% recognition accuracy results. KarsligEi at el. 1 designed and implemented a new semi-supervised anomaly detection system using the k-mean algorithm. By using the k-mean algorithm firstly, they separated normal samples to clusters. Then, they calculated a threshold value to distinguish normal and abnormal samples. For their study, they utilized NSL-KS a labeled dataset to find out the effectiveness of the designed anomaly detection system. In their experimental results, they showed that their designed and implemented anomaly detection system is able to achieve 80.11% accuracy results by using NSL-KDD data set. While Ahmed [14] addressed the detection problem in network and show studied that identification is an very significant task in network traffic classification that it identifies and detect the anomalous data from the provide dataset. They also highlights that intrusion detection is an essential and an exciting area that has been widely used in machine learning and statistic.

Supervised learning: Supervised learning method is a ML technique also called classification or identification technique. In this method, a labeled dataset needs to identify the unknown classes. In this technique the Input data is known training data testing data. In figure 4 are shown the supervised learning method in which shown that first trains the model with some labeled data and then supervised learning method used the model with new input sample data to make

prediction output. This technique infers function in the labeled training data and start TS with training dataset. In the equation, one xi indicate the feature vector and yi is the output mean predicted results. However, example problems are regression and classification. While examples algorithms are exist on Logistic Regression and Back Propagation Neural Network.

$$TS =< x_1, y_1 >, < x_2, y_2 >, ... < x_n, y_n > \tag{1}$$

Unsupervised ML: This technique is known as cluster technique. In unsupervised learning technique, there is no need to label the data, or there is no need to complete labeled dataset. The unsupervised method is a type of machine learning technique. Thus this technique doesn't need any labeled or predefined dataset for learning to classify the unknown classes. Examples problems are clustering and dimensionality reduction, and association rule learning. While the example includes the Apriority algorithm and K-mean algorithm.

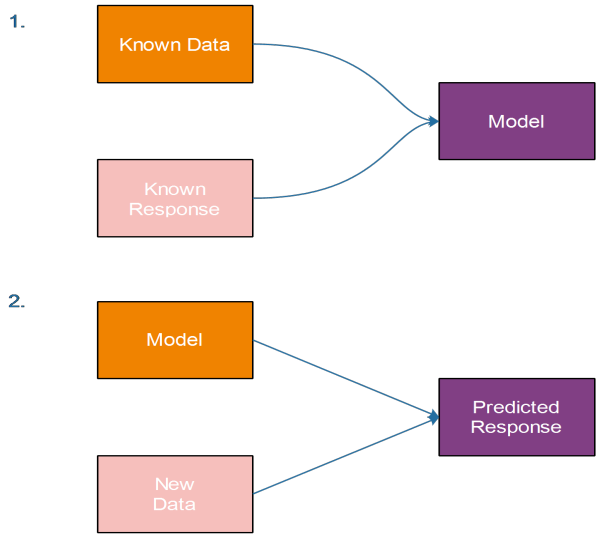

Figure 4: SUPERVISED LEARNING METHOD DESCRIPTION

5.1. Artificial Neural Networks

Artificial neural network (ANN) is a computational which is based on biological neural networks (Structure and Functions). ANN is also called a neural network. The main structure of the artificial neural network can be affects, if the information is flowing in the network. Because the neural network learning technique based on input and output layers. ANN is nonlinear statistical data modeling tools where the multipart relationship among the input and output data are model. However, ANNs are interconnected artificial neurons able for several computations process on their inputs. ANN algorithm become very popular when Support Vector Machine (SVM) was invented. Basically, ANN classifiers based on the perception. However, ANN and SVM take long running time during learning times such as in ANN, when the numbers of feature more the learning running time of ANN increase with respective. Similarly, when the number of feature less the running time will be less. It is also the capabilities of ANN classifier to generate nonlinear models with one or more hidden layers.

For accurate network traffic classification, QoS control, network security and many other network activities in [70] a new method was proposed feed-forward neural network for effective Internet traffic classification, used to separate the drawbacks of port-based and payload-based identification techniques. However, after a lot of experiments, a comparison conducted to clarify the exploration of the new produced method. They found in their research work that fast correlation technique for feature selection gives efficient performance results through neural network technique as compared with other machine learning methods. In their study, they used dataset which is originally developed by Moor el [64], and they conduct analysis on the results of new method. For better performance evaluation, they used 0% to 100% scale for the accuracy result. Runyuan Sun et al. [71] developed a new technique for accurate traffic sample collection named DHTCP and they showed that their develop platform is able to collect sample on user host with application information. They used a probabilistic neural network using (DHTCP) trace dataset Internet traffic classification. And then they used a set of statistical features as describe earlier to describe the traffics.

For the evaluation of the performances of the RBFNN, FNT and SVM TPR, FPR and accuracy matrices are employed [72, 73]. In this study, the most predominant Internet traffic Web and P2P traffics are studied. In their research study, they clarify that the PNN is an efficient ML technique for network traffic classification. Bivens et al. [74] demonstrate the complete Intrusion Detection System (IDS). However their study focus on the detail IDS system such as they cluster the normal traffic, preprocessing stage as well as ANN decision and normalization[75]. In this study the initial stage used a type of unsupervised Artificial Neural Network (ANN). It is also known as a Self-Organizing Map (SOM) and used to learn the traffic over time such as well-known TCP/IP port numbers. In this method, initially the features set are quantized into bins and then the quantized bins fed in the coming stage, an ANN and then in the first stage the numbers nodes and as well as layers are determined. However after the completion of MLP training the identification of intrusion in a traffic can be started in a network. More in depth, for learning new traffic pattern the system can be restarted and ML algorithms can be trained. However, they reported successfully for prediction of regular behavior. The proposed technique is effective and efficient for intrusion detection. However, some attacks were not effective detected which reached up to 76%.

5.2. Bayesian Network

It is a model that describe the used variable and their relationship which is also known Probabilistic Graphical Model (GMs)[76, 77]. In this method the network is designed with the node of variables which are the discrete or continuous and while their directed edges shows the connections between them. Similarly, in the network the child nodes are reliant on their parent's node [18]. Every node keeps up the positions of the random variable and conditional probability form. Considering Internet traffic identification, Lavadas et al. [78] compare several ML techniques for the classification of C2 flow of IRC-based botnets traffic. They divide the whole task into two stages. In the first stage their study exist on charactering the IRC and Non-IRC traffic while the second

stage is differentiating botnet and IRC traffic in a network [79]. They used in their study a collection of 18 locations TCP-level data at Dartmouth University Computer Campus with a duration of four months. However, the TCP data are utilized to generate the NetFlow data or the network streams data. They used the filter layer for the extraction of IRC in all network data. However, in first stage, a Naive Bayes classifier gives an effective performance (2.49%) for low FNR and (15.04%) low FPR for the flow of Dartmouth traffic. But the results of the Bayesian network is with respective precision and FNR. While using other machine learning classifiers gives 97% precision. But the False Positive rates results are greater. Auld T et al. [80] present a new ML classifier for the network traffic classification which achieves very efficient accuracy results without using any information of application such as port information etc. They used supervised machine learning based on a Bayesian trained neural network. They used derived feature from the packet content for their technique. Similarly, the training and testing steps are also applied on these derived attributes from the packet stream content of packet head [81]. They collected the data by high-performance network monitor which are effective for network traffic classification. However, the training and testing sets are limited. These traffic flows consist of only TCP and with complete TCP connections. However, the complete flows are flows in which the entire setup and tear-down are viewed. Although, they are also compared in their study Bayesian neural and naive Bayesian estimator. However, they showed that Bayesian trained neural network is capability to identify network traffic, which are based on derived statistics features and no port or host identification information and the accuracy improved from 50% to 70%. Finally, they showed that a small number of a derived feature of packet content carry enough information for Internet traffic classification. Which are very useful for machine learning traffic classification.

Kruegel et al. [82] presents a novel method for the performance of Bayesian identification event for intrusion detection [83, 84]. They improved the naive threshold based schemes traditionally used by Bayesian networks. In thier study they used DARPA 1999 data set by TCP/IP packets and a set of attribute of

call are used for the study. However, the attributes are used in the Bayesian Network for the calculation of the probability of a normal state. Different threshold values are used and achieved different values such as 75% accuracy, 0.2% FAR. Similarly, 100% accuracy and 0.1% FAR with different threshold values are reported.

5.3. Decision Trees

This method tree is a type of supervised learning algorithms having a well-defined variable that is used for a specific problem classification. Decision tree can be used for categorical and as well as continuous input and output variables. In more depth, a decision tree is just like a tree that has leaves, known braches you may say to represent classification. Which represent a combination of features that leads to identifications. To build the decision tree model automatically there are two methods which are very popular such as ID3[85] and C4.5 [86] machine learning classifiers. These two algorithms are design on DT which used training set with the concept of entropy knowledge.

As we discuss in the above lines that decision trees are two types Categorical Variable Decision Tree (CVDT) and another one is Continuous Variable Decision Tree (CVDT). It's mean that a DT which has categorical target is known as CDT and a decision tree which has a continuous target variable then it's know as continuous DT. The terminology connected to decision trees are Root Node, Splitting, Decision Node, Leaf or Terminal Node, Pruning, Bruch or Sub-Tree and Parent and Child Node which are the basic terms usually employ for DT. The DT is very easy to understand; even nonanalytical background students can easily understand the output of the decision tree because its graphical representation is straightforward to understand and a user can relate the hypothesis. A decision tree is the one of the speedy way to find out the relationship between variables and create some new variable which is effective for the identification of the target variable. Compare to other modeling techniques decision tree need less data. The most essential thing in this technique: it can utilize both numerical and categorical variables.

In our previous study [87], we study the effective packet numbers for early stage Internet traffic identification using ML classifiers. In which, we study to identify the best packet numbers. For this aim, five network traffic datasets are conducted. Twenty packets sized are extracted, and then mutual information (MI) analyses are conducted to identify the mutual relationship between the packets. After extraction ten well-known machine learning algorithms are utilized using crossover identification technique. In more depth, to identify whether the selected effective packet numbers are really useful or effective two statistical tests known as Friedman and Wilcoxon test are conducted. Which show that the selected packets numbers are really useful at early-stage network traffic identification. However, in more depth, the conducted ML classifiers results are very efficient compared with each other's to find out effective machine learning algorithms. In which decision tree C4.5 machine learning algorithm gives a very useful performance. In this papers the authors used two different datasets at early stage network traffic identification, in both datasets comparing with other ML classifiers, the decision tree machine learning classifiers gives outmost performance results. Similarly to address the problem of effective features selection for network traffic classification in [88] a features selection called Weighted Mutual Information (WMI) is introduced. After that, a feature selection algorithm called WMI_ACC is proposed which filter the features with accuracy. To select and identify the powerful attributes for Internet traffic identification well-known five ML classifiers are conducted to show that the proposed algorithm performance results. However, two different network environment datasets are used for this study and to identify the effective attribute set from the selected set of attributes. Their study results shows that the proposed method is best for feature selection and the applied five well-known machine learning classifiers gives very promising results but decision tree C4.5 machine learning algorithm give very promising performance as compared to others conducted ML algorithms. It's mean that in network traffic identification decision tree ML algorithm is very useful to classify Internet traffic accurately.

5.4. Ensemble Learning

It is a machine learning method in which many learners are trained to solve a specific problem combine. In other words, an ensemble includes multiple numbers of base learners called learners. The generalization of an ensemble is much robust as compare to base learners. The base learner is also known as "weak learners". But the most theoretical analysis does work on feeble learners. Ensemble technique combines multiple hypotheses to make a good one compared to the best hypothesis. Sometime ensemble technique execute multiple weak learners to make a strong learner [18]. In [89] the authors studied that "Boosting" is a technique used to improve the performance of any machine learning algorithm [90]. To improve the performance and remove the error rate of any weak learner a booting can be utilized that need a random guessing. However Adaptive Booting means AdaBoost is one of the most and most used technique by researches used to identify and remove the of noise an inherent to machine learning. Similarly, bagging (bootstrap aggregating) is also a useful method used to reduce the over-fitting like Adaboost ML algorithm and improve the generality of the predictive model. Similarly, the Random Forest classifier is a ML technique that used to combines ensemble learning and decision trees. Using trees the forest takes features as input randomly. Zhang et al. [91] studied and outlined three different DM based models for network ID. For the misuse, anomaly and in intrusion detection, random forests algorithm are used. For the rule-based system problem, a random forest algorithm is conducted to model patterns of intrusions. With the help of Java programming in WEKA environments[45] and Fortran program [92] the proposed technique are conducted. For the proposed technique different datasets are achieved from the KDD'99 datasets. However, for the misuse framework, the offline phase pattern of intrusion are model. While the model can quickly find out in real time using the built patterns. For better performance and better accuracy results feature selection algorithm are used of the random forest algorithm as well as sampling methods are also used for better performance. Similarly, in their study, they used outliers identification of the random forest classifier in anomaly identifi-

cation. Because random forest ML algorithm is a supervised machine learning classifier and used label dataset to shape a model. Thus, their technique develop a pattern of Internet services. In this pattern, the technique identified the outliers which is related to the patterns. However, their proposed technique divide the dependency on malicious free training data, which was a very big and fundamental problem of supervised intrusion detections. Moreover, their introduced approach achieved effective performance results.

5.5. Naive Bayes

Nave Bayes machine learning classifiers are generally probabilistic machine learning classifiers using Bayes theorem [93]. Naive Bayes is a classification technique rely on Bayes Theorem. In simple words, a Naive Bayes suppose that the existence of two specific feature in a class and the one feature unrelated to the other features. However, the Naive Bayes technique is not difficult to build and very appropriate particularly for a big dataset. Moreover, Naive Bayes is also very effective for the highly complicated sophisticated identification or classification methods.

Naive Bayes classifier also has several disadvantages. This classifier an useful mean if the attributes are conditionally free given in the correct class. However, Naive Bayes machine learning classifier is the one classifier that mostly users like due to understanding the model easily as compare to other ML classifiers as SVM. In our study [46] Internet traffic classification technique. In this study technique of a internet traffic identification is discussed using ML technique and show how new researcher will conduct and utilized the network traffic classification. As this study is about the comparison of machine learning techniques to identify the best one algorithm out of four ML algorithm. However, the experimental result shows that Naive Bayes ML algorithm also effective for Internet traffic identification. In Internet traffic identification and network management in [94] WeChat different services traffic are classified using particular network datasets. For accurate text messages flow service traffic identification using Wireshark tool two datasets are build and then from trace

traffic 50 feature are extracted. After extraction features, four machine learning algorithms are used to for classification. However, from the results analysis, the used ML algorithm performance are efficient, but Naive Bayes gives little bet low-level result as compare to other ML algorithms. In [95] the author extends the study of WeChat text messages service traffic identification [94] to text and picture messages service flow traffic identification using machine learning algorithms. In this study, authors used the same method but just added WeChat pictures services with text messages and used four ML algorithms for traffic classification. In this research work, Naive Bayes achieve effective results as compared to Bayes Net ML algorithm. In [87], for the effective packet numbers at early stage Internet traffic identification using ML classifiers. In which, we study to identify the robust packet numbers. For this aim, five internet traffic datasets are conducted. Twenty packets sized are extracted, and then mutual information (MI) analyses are conducted to identify the mutual connection between the packets. After extraction ten well-known machine learning algorithms are utilized using crossover identification technique. In more depth, to identify whether the selected effective packet numbers are really useful or effective Friedman and Wilcoxon statistical tests are conducted. Which show that the selected packets numbers are really useful at early-stage network traffic identification. However, in more depth, the conducted ML algorithms performance are better compared with each other's to find out effective machine learning algorithms. In which decision tree Naive Bayes ML algorithm achieve effective performance. In this paper, the authors used two different datasets at early stage network traffic identification, in both datasets comparing with other machine learning algorithms, the Naive Bayes machine learning classifiers gives outmost performance results. Similarly to address the issue of efficient features selection for Internet traffic identification in [88] a features selection WMI is introduced. After that, attributes selection algorithms called WMI_ACC is proposed to filter the attributes with accuracy. To select and identify the effective features for Internet traffic identification well-known five ML classifiers are conducted to clarify that the introduced algorithms results are efficient. However, two differ-

ent network environment datasets are used for this study and to identify the effective attributes from the selected attributes. Their results shows that the proposed method is best for feature selection and the applied five well-known ML algorithms give very promising performance, but Naive Bayes machine learning algorithm gives very promising performance as compared to others conducted ML algorithms. However, it clear that Naive Bayes machine learning classifier is useful for Internet traffic classification.

5.6. Genetic Algorithms

Genetic algorithms (GA) is also known as well-regarded evolutionary algorithm in the history, or we can say that it is an algorithm which is based on principles of selection and genetics in [96]. Similarly, in [97] the author discussed GA the main concepts and terms of the algorithm and also gives very effective improvements in the algorithm such as mutation, crossover, and selection. They also studied and investigate the main application in the field of image processing. In the end, in his research work, they showed and summarized that GA algorithm gives effective results and did very efficient and build identification images when there is several pixels and very large. The main point that they find out that the numbers of chromosomes and generations are the leading main roles. Others parameters such as mutations etc. are also examine in the research work. Similarly, in 2019 Mishra at et [98] discussed the GA algorithm with detail such as they have mentioned that GA algorithm includes initialization, mutation, crossover, and selection. Similarly, in 2019 the author in [99] discuss and proposed a genetic algorithm known GA- OCSTuM. They used the proposed technique for anomaly detection for huge sensor data in the internet of things (IoT). In experimental result analysis, they clarify that the introduced technique is able to increase the accuracy and efficiency of anomaly identification in data. However, in terms of the intrusion detection system the three very important things are speed, adaptability, and accuracy. Though, Genetic algorithm produced very promising performance compared with the rest of ML algorithms in the intrusion detection system for effective features selection. However, there

is no surety that the algorithm will identify the global optimum. Furthermore, describing problem space in the classifier is very complicated. They required a huge fitness function analysis [98].

5.7. K-means Clustering

It is very popular algorithm in unsupervised learning algorithms. This is used to recognized and find the unlabeled data in different clusters. However, for the implementation of this algorithm two parameters are needed, the dataset and the numbers of clusters. To resolve clustering problem and if If the desired number of cluster is k, then by using k-mean clustering algorithm the first step is to initialize k cluster, then using distance function every node with the nearest centroid, and then assigned new centroids according to the current node and then stop the classifier, if convergences is true, then return to second step again [100]. However, an anomaly detection system. Many researchers used k-mean clustering algorithm as a classifier find out anomaly and normal traffic or data instances, and some used the algorithm as a compaction method to separate outliers and training set and produce refined data set for the machine learning algorithm. However, both of them technique are effective for intrusion detection systems. In [98] mentioned that this method for anomalies detection fails if the noise in data build clusters by themselves[98]. Thus on this case, the algorithm will be not effective for anomaly detection or to identify intrusions. For anomaly identification the authors in [101] want to find out the and solved this problem bu using the two-step technique. They firstly modified the k-mean classifier and used for break down the data pattern or objects. While in second step they construct a max heap which is depend on the points number in the cluster. For their research study, they used iris dataset. From the analysis of their experimental results, they showed that the proposed technique effective identification results. However, they only used the approach on iris dataset and mentioned that the method could be applied and implement on social media intrusion identification based on comments in the messages. Likewise, for the detection of an outlier in [102] the authors interested to use multi-dimension

data, like multi-dimensions, No of fraud cased, etc. However, they proposed a new technique for detection outlier base on PSO with enhanced K-mean algorithm using air quality time-series dataset. They showed that the proposed method is effective for the detection of outlier in air dataset. They also discussed that the proposed method can be applied in the different dataset such as anomalies detection and network traffic classification, health care, etc.

5.8. Reinforcement learning

It is nowadays going popular and hot topic. A type of machine learning technique involve an agent, and an action space. The agent enable to learn in interactive environment. An agent takes action. For instance: In real life example an agent is me and you. Now, what is action? Action can be a set of moves. For example, in a video game moves list, moves can be left or right while the discount factor is multiplied by future awards. Similarly, Environment can be discussed as it takes the state of the agent and action as input. Whereas, State can be, a concerned and instant condition in which the agent finds itself. While reward is the agent feedback such as the measurement of an agent success or failure action and policy is the strategy. As we discussed, RL is a very interesting area in ML, where more then one agents and ML collaborate with each other for learning behavior to increase the performance of the attacks. In [98] the author studied that agents become aware the environment and input is outlined to locate the exact information. However, when the reinforcement learning executes the action and feedback is conducted in the environment. They also mentioned that right actions of agents are gives reward by the environment, which is known reinforcement signal. And after that the agent leverage the rewards and goes to increase the learning about the environment to take the next steps. However, for anomaly detection, some researcher has applied the Reinforcement Learning (RL) to identify the malicious attacks [103]. The author in 2019 [100] discussed the RA and shows that using RL, the agent can be known as controller and the network can be say environment. In their study, network status is monitor by the controller and learns some decision for the controlling

the data forwarding. In the diagram st is the monitors a state and at is time step. Similarly, st select action from the available action space A receive and quickly gives reward rt which shows the status of action as good or bad status is, and then take next step st+1. Here main responsibility of agent is to learn behavior policy alpa and then the agent can take appropriate action in a specific state[100].

5.9. Deep Reinforcement Learning (DRL)

The basic advantages of DRL method is that the algorithm work without prior to an exact mathematical model of the environment. DRL uses reinforcement learning, and deep learning principle in order to create effective machine learning algorithms that could be applied in any area of computer science. DRL is capable to solve complicated problem and decision making task in the area of the video game, health and computer science, etc. in [104]. Deep RL is able to solve the problem that was no solved previously. Deep RL has some new achievement are in [104], attaining superhuman-level performance in playing Atari games from the pixels Mnih et al. in [105], mastering the game of Go in [106] etc. Though, every learning approach have some advantages but have some disadvantages. Here, RL has some disadvantages, like low convergence rate and its inability to solve high dimensional data. These disadvantages can be define by [100]. Similarly, in a [104] study that, the above-discussed achievement is significant for popular games, because of the variety of complex and diverse tasks that need with high dimensional inputs, 2018, in [107]. In fact, Deep RL system are now in use. For instance, Gauci et al. in 2019 in [108] define, how Facebook utilized DRL systems such as for pushing notifications and faster video loading in smart prefetching.

5.10. Reinforcement Learning Based Game Theory

It is basically related mathematical modeling, which focus on strategic rational and decision makers. For instance, as we know, a game includes a number of players, a collection of policies and a number of package functions. In this

technique, players are known decision makers, and players used the policy or utility functions to select ideal strategies. In-depth, there are two sections of game theory, which are: Cooperative game theory (CGT) and non-cooperative game theory (NCGT). In (CGT) games, mostly players cooperates and form multiples coalitions. We can say, that a group of player called coalitions. In this theory, this is the primary units of decision making and enforce to cooperate behavior, and this system can be seen between coalitions of players. While, non-cooperative game theory, the players compete with each other against and choose a strategy for their own utility increment [100]. Mostly researchers used non-cooperative game theory in the field of network. Moreover, at the beginning of the game, players don't communicate with each other and don't know any strategy information about each other. While the set of all players shows at the end of play round their certain strategies. In this reason, every one player utility can lead to be effected by other player utility. However, it is important to used adoptive learning method to identify the strategies of the other set players. Reinforcement Learning is a widely used system, which can help the player in the game to select optimum strategies by using historical information. Therefore, RL game theory is an efficient decision making theory for decision making.

5.11. Fuzzy Logic

Fuzzy logic is an approach that computes the degree of truth, or another word the method is based on the degree of truth mean true or false. However, this technique firstly introduced in 1960s by Dr. Lotfi Zadeh at University of Califonia. Fuzzy logic can interpret the properties of a neural network. Neuro-fuzzy is very popular in the area of Intrusion Detection. The authors in [98] discussed that the fuzzy logic is not detect all types of attacks. However, the performance of this algorithm become very well, when it applied with other machine learning classifiers. Fuzzy logic has also been used in correlation with intrusion detection systems. However, the primary key characteristics of the fuzzy logic are [109]: 1. The fuzzy rules allow constructing the if-then rules and on security application based can be modified. 2. Can combine the input from

varying sources. 3. The quantitative measures used by IDS are fuzzy in nature, such as CPU usage time and connection interval. 4. Often the degree of alert of IDS fuzzy. However, the author in [98] point out some disadvantages of fuzzy logic such as they mentioned that the fuzzy logic system need more fine tuning and simulation before operational. Second, they discuss that it is challenging to develop a model from fuzzy systems with comparing other machine learning solutions, because of complexity in the fuzzy model.

5.12. Support Vector Machine

Support Vector Machine (SVM) is also known SMO. SVM is known as a supervised machine learning technique. SVM is widely used in many areas for identification and classification. It is very useful for both classification and regression, which formally defined by a separating hyperplane. In other terms, the labeled data (supervised learning), the output of the algorithm is the optimal hyperplane which classify new examples. SVMs are famous for generalization and also very effective when the number of features is high, and numbers of data points are low. In [18] studied that when two classes are not separable, slack variables are added and a cost parameter is assigned for the overlapping data points. They also studied that with quadratic optimization with a practical runtime a maximum margin and place of the hyper plane is determined. In the work [110] the authors proposed an approach for the evaluation of NetFlow records by referring to a method of temporal applied to Machine Learning techniques. They used NetFlow data collected from the real world using Flame tool and more others Internet Service Provider Services such as NeBIOS scans, DoS attach, POP spams and Secure Shell (SSH) scans. However, the study working a one-class Support Vector Machine classifier. In this study, they propose a new kernel for anomaly detection based on the time position of the NetFlow data. Experimental results show that the performance of different types of attacks reported as 89% to 94% accurate on attacks with FP rates of 0% to 3%. However, for precise network traffic classification and identification according to application types work in [111] using Support Vector Machine (SVM) a

classifier is presented. The classifier is proposed for specially P2P application traffic classification with statistical characteristic of network traffic. However, P2P, BitTorrent, PPLive, Skype, and MSN traffic were considered. They presented in the study that how to achieve and label the traffic samples, describe the attributes of traffic selection methods. The results show that the proposed approach is effective for encrypted traffic classification and good to identify network flows of application layers.

Accurate network traffic classification is an important technique for network traffic management, anomaly detection, network security, Quality of Service management and more. In the last decade network traffic classification got very importance due to rise up day to day internet users traffic. It is very important for Internet Service Providers (ISPs) to classify internet traffic accurately. For this purpose of network traffic classification, researchers proposed many effective classification models to classify internet traffic accurately. Traditional technique such as Port-based and Payload-based technique as we discussed are not very effective due to some limitation. Then, approach machine learning based technique is proposed to overcome these limitations. However, it is also important to discussed every aspect related to network traffic classification such as network traffic classification at early stage. To the best our knowledge, we are the first to discussed this topic in survey paper. In section 5.13 the details information are given.

5.13. ML And DM Methods for Early Stage Internet Traffic Classification

Recently early-stage network traffic classification is proposed in network traffic classification, and many researchers contribute in early-stage internet traffic classification. However, there is no survey study on early stage internet traffic classification to the best our knowledge and this is the first survey paper for early-stage internet traffic classification and Data Mining and Machine Learning Methods for Internet Traffic Classification. In this section, we survey eleven papers existing approaches for early-stage internet traffic classification and related topic that were published during 2011-2017. In this section, we have tried to dis-

cuss every aspect of network traffic classification. Network traffic classification or network traffic identification is the most crucial technique for network traffic management, anomaly detection, network security, Quality of Services management and more as the author in [83] discuss that the network traffic classification is the method to identify internet application or protocol that exists in a computer network. From the perspective of performance management, it is very important for both Internet Service Providers (ISPs) and network operator to manage the overall performance accurately. Internet traffic classification plays a very significant role in network security and network management, such as Anomaly Detection, Intrusion Detection, and Quality of Service (QoS) management. Moreover nowadays internet traffic growing [16] day to day and consume large size of bandwidth such as Peer-to-Peer application traffic consume large size of bandwidth on the internet. Why these P2P applications consume large size of bandwidth because its traffic flows are involved in both directions and at the same time. Thus from the management point of view P2P consume a large amount of bandwidth. Moreover, not only the P2P applications consume large size bandwidth of the available bandwidth as well as other kinds of network traffic such as SMTP, FTP, HTTP, etc. also take their share of available bandwidth. It means that the ISP is able to implement Quality of Service (QoS) policies for each application. Similarly, ISP facing many problems like satisfying costumer with effective quality broadband experience as well as buying expensive backbone links and upstream bandwidth. So internet traffic growth is increasing by proposing various types of internet application like P2P, HTTP, SMTP, FTP, etc. but the major contribution in content sharing or content distributing in these applications is P2P application like audio, video, and game which possess too large size [84]. Nowadays, mostly P2P application traffic possesses on 60% of total internet traffic [18, 19, 20], and which hold mostly portion of network bandwidth. Azzouna and Guillemin [22] in their study they showed that 49% traffic are generated due to Asymmetric Digital Subscriber Line (ADSL) link application. Similarly, Ipoque [39] conducted worldwide study Internet traffic in 2007 and showed that mostly P2P file sharing applications produce traffic

compared to other Internet application traffic [24]. So the classification or identification of the application that provides more and more traffic is very difficult to manage Quality of Services, implementing security and intrusion detection in each flow application and this task is very crucial for Internet Service Providers (ISPs). However, traditional internet traffic classification completes the internet traffic classification task which includes on port-numbers of transport layer application protocols. But this method of application identification failed due to introducing various applications of using the random port number for data transformation. Moreover, some other application starting the masquerading technique by using a well-known port number to hide their traffic such as port number 80 used by HTTP. Madhukar A [29] showed in their study that internet traffic classification is very crucial to effectively identify internet traffic by using port-based number techniques. Due to using many applications random port numbers this method failed for the identification of internet traffic. Another method Payload-based technique is proposed for the classification of internet traffic. Although this technique for traffic classification is very effective and achieves high identification performance, but this method was some limitation such as need of expensive hardware, computational resources, privacy issues as well as another limitation of this method is that this technique cannot be for encrypted data application traffic etc. so this technique also failed due to so many limitations. Overcome the limitation of payload based traffic classification technique. Another method statistical or behavioral based method was proposed. This technique total number packet send, packet length, packet size, packet received, etc. this technique does not possess the limitation of port-based and payload based technique. Through these techniques, Internet Service Provider (ISP) or Network Operator can take some action such as resources management, block some flows which are not authorized flows. With internet traffic classification the development growth of internet application can find for managing its traffic accurately. The main objective of this survey paper is to deliver a comprehensive overview of numerous traditional method as well as the existing ones for early-stage internet traffic classification. There are many sur-

vey studies on Internet traffic classification [85, 86] etc. in these survey paper all the aspects of traffic classification are discussed such as method s for encrypted traffic classification and analysis, machine learning traffic classification, etc. but no survey or review study has been proposed on Early Internet Traffic Classification. This survey study about early stage internet traffic classification is the first survey study.

Recently Internet traffic classification or identification got very pivotal importance in scientific contribution due to a bundle of issues related to internet traffic classification, such as managing network traffic accurately, providing network security, maintaining quality of service for application, anomaly detections, etc. As we discussed that there is no survey study has been proposed on early stage internet traffic classification. This is the first survey study in which early-stage Internet traffic classification are also discussed. However, from the last six years, early-stage network traffic classification got a very pivotal interest in the research community. Using machine learning technique most researcher extract features on a whole traffic instances [112, 113, 114] . A. W. et al. extract 248 statistical features based on the whole traffic. Like packet size and minimum and maximum as well as average values. While using these statistical features, all the machine learning classifiers can get high accuracy results in network traffic classification.

However in real circumstances its identification result is not very effective to recognize the internet traffic applications. Thus it is very important to identify internet traffic accurately. For this purpose, few researchers have tried to build an effective model for network traffic classification using early-stage network traffic. And then this early stage network traffic classification becomes a hot topic in network traffic classification [115]. Early stage traffic classification means that the classification of internet traffic with early few packets of flow with statistical features. In 2006 L. Beranille et a. [66] use the size of first few packets of each TCP flow as the features and apply K-mean clustering machine learning technique using ten ten types of application traffic. They got very high identification result. The results that they got in their study open new possi-

bilities areas for online internet traffic classification. They used the first five packets of a TCP connection for the identification of application. After that in 2009 A. Este et al. [116] proved that early stage packet of internet traffic carry enough information for traffic classification as well as they analyzed that the packet size is the most important and effective feature for early-stage network traffic classification. They only focus on TCP traffic. While the three datasets UNIBS, Auckland and third dataset that they used for their study is Internet Exchange Point NZIX data set. In this survey paper, we focus on early-stage traffic classification and its related technique and approaches for internet traffic classification as well as and its limitations to better understand also internet traffic classification. However, it is also important to give the touch more deeply related to anomaly traffic identification using machine learning algorithms.

In this section, we have tried to introduce all the important aspects that are very valuable in Internet anomaly traffic classification not only ML-based anomaly detection. So, we have tried our best to cover all the aspects related to anomaly detection.

Though, many researchers focused on anomaly detection in different and proposed effective anomaly detection systems and models, but to best our knowledge no one proposed early stage anomaly detection system, so on this reason. We review papers related to early-stage internet traffic classification. We start the selected papers from 2011-2018 and year by year we review the selected paper related to early-stage internet traffic classification means first we review the papers that are published in 2011, and then we review the papers that are published in 2012 and up to 2017. Notice that we are only interested to review those papers which were only related to early traffic classification in section.

1. Early Identification of Peer-To-Peer Traffic The work published in 2011 by Hullr et al. [117] analyze the first few bytes of a few packets of each flow of P2P TV and file sharing applications. They extend and improve the work publish in [118].They proposed an automatic machine learning approach, which got quite good results for early traffic classification. They showed

that P2P application can be classify using only the first flow few bytes of a few packets. They used limited dataset as the first 16 bytes of the first packet each flow and got very high accuracy to result up to 95%. But this method only classifies few traffic application protocol types as well as this approach requires less computational and memory resources compare to DPIs with performance. Moreover, another advantage of this approach is that there no need for human expertise developing pattern (training) for method. Additionally, they capture the traffic on their campus and build the dataset and used Random Forest machine learning algorithm. For the performance evaluation of the traffic classification method the authors used True Positive (TP) and False Positive (FP). Moreover, the authors used two most popular DPI tools in their research paper OpenDPI [119] and Tstat [120] as well as payload based method Coral [7] for ground truth.

2. Early Classification of Network Traffic Through Multi-Classification The work by Dainotti et al. in [121] published in 2011 present high effective classifiers and applies a hybrid feature extraction method for early-stage traffic classification. They proposed and evaluated the combination techniques for internet traffic classification including BKS-based algorithms as well as they propose for the first time use of multi-classification for early stage traffic classification. While in this study the author means from multi-classification, the approach that combines multiple classification techniques through specific algorithms to build useful and accurate multi-classifier. They say in the paper that the overall accuracy can improve by the combination of stand-alone classifiers that show complementarities, as well as Behavior Knowledge Space combiners, is more effective than other traffic classification. They also write in his paper that the literature review on transport-level port is a useful classification feature, but combiner cannot efficiently exploit the power of a port-based traffic classifier. Moreover, they used PS, IPT, Average Standard Deviation, L4 Protocol, Biflow duration and size, IPT statistics, Payload; Ports features

conducting J48, K-NN, R-TR, RIP, MLP, NBAY, PL, PORT machine learning algorithms and for combination they used MV, D-D, BKS, and WER.

Table 3: Different Features Dataset Results

Features Set	Accuracy			AUC		
	Auckland	UNIBS	UJN	Auckland	UNIBS	UJN
PS	0.9880	0.9879	94.06	0.9954	0.9963	0.9331
Diff1	0.9856	0.9561	0.9397	0.9947	0.9960	0.9180
Diff2	0.9859	0.9810	0.9413	0.9921	0.9939	0.9167
Statistical	0.9871	0.9603	0.9166	0.9944	0.9796	0.9231
MI	0. 9890	0.9870	0.9429	0.9960	0.9957	0.9299

3. Issues and Future Directions in Traffic Classification In 2012 Dainotti et al. [115] review achievements in internet traffic classification and discuss futures directions in internet traffic classification with their trade-offs in reliability, privacy and applicability. They outline unsolved challenges in the field of network traffic classification and suggest several strategies solving these challenges. According to Dainotti extracting features set from few packets can gives many benefits like lower computational complexity for features extractions. They outline that are the extracted simple features from few packets containing enough identification? The fundamental problem of early stage internet traffic identification is to find out the most effective features for early stage network traffic classification.
4. Timely and Continuous Machine-Learning-Based Classification for Interactive IP Traffic Nguyen et al. [122] in 2012 derived statistical features from sub-flows for VoIP traffics identification. Whereas sub-flows mean a small numbers of most recent packets taken at any point in flows and

they extend early stage concept to timely. Where has novel approach showed that using sub-flow size 25 packets the Nave Bayes machine learning classifiers gives 98.9% recall and 89 precision classifying ET traffic. While classifying VoIP traffic using Nave Bayes classifier gives 99.6% recall and 97% precision results values. Moreover applying C4.5 Decision Tree machine learning classifier perform 99.3% recall and 95.7% precision for classifying ET traffic and classifying VoIP traffic they got 95.7% recall and 99.2% precision results. Note that they used Packet lengths and packets inter arrival times for this study work.

5. On Accuracy of Early Traffic Classification In 2012 Qu at al. [123] have studied accuracy issues of early stage network traffic and found that it is possible to identify accuracy with effective classification traffic. In this paper they used packet size, inter arrival time and direction extracted at the flow level of first few packets using UNIBS 2009 anonym zed Internet Trace and UNIBS 2009 SSH tunnel traces data sets they applied three machine learning classifiers C4.5 Decision Tree, Nave Bayes and Support Vector Machine using two publically available datasets.

$$diff1_i = ps_{i+1} - ps_i = 1, 2, ...9. \tag{2}$$

6. Application Traffic Classification at The Early Stage By Characterizing Application Rounds

   In 2013 Huang, N at al. [124] proposed machine learning based high accuracy algorithm Application Round method (APPR) for the identification network application traffic at early stage. By analyzing the behaviors of different application they extract features of early stage traffic. For some classifiers they used packet size inter packet time of the first ten packets and average and standard deviation values of packet size and inter packet time of the early packet for other machine learning classifiers. By using these features and applying machine learning classifiers they got achieve high performance results. They focused on 59 protocols to proof the proposed classifier and also to compare the accuracy level with other

machine learning classifiers. By applying J48, PART and Bayesian network algorithms, they achieved very high accuracy results by comparing other machine learning method with the same trace traffic while for normal ratio his method achieve 7-8% improvement and for fixed ratio has method got 15-30% improvement of overall accuracy.

7. Low Complexity, High Performance Neuro-Fuzzy System for Internet Traffic Flows Early Classification. Rizzi et al. [125] in 2013 proposed Neuro-Fuzzy system for early internet traffic classification based on simple features extracted from first few packets using two different data sets. They showed in has paper that Min-Max model trained by PARC algorithm can achieve accuracy close to SVM model. While they extract Flow direction, length, and timestamp features for early traffic Classification using PARC algorithm and SVM machine learning algorithms.

8. Features Evaluation for Early Stage Internet Traffic Classification The most interesting work published by Peng L et al. [126] in 2014. In this paper work they focus effective difference between packet size and its statistical features for early stage traffic classification. They evaluate the features for early stage internet traffic classification using mutual information analysis. For features evaluation they extract first ten packet packets sizes and its derived statistical features on three traffic data sets (Auckland II, UNIBS and UJN traces) as well as they also find in this paper first and second order difference in features. By the help of equation 1 they find the first and second order difference. The Auckland II and UNIBS data sets are publically available dataset while UJN data set is his own trace set. Then they execute seven well-known machine learning classifiers. They showed in his paper that derived features such as difference and the statistical feature are not very effective features to compare to original packet size features and shows that features select through mutual information analysis method achieve very attractive identification accuracy results in most cases. Below Table 5 show the Accuracy and AUC result using original feature set, difference, statistical and combine features data set.

9. Effectiveness of Statistical Features for Early Stage Internet Traffic Identification. In 2015 Peng L et al. [58] work on early stage traffic to find out the effectiveness of statistical for early stage internet traffic classification. In this paper they evaluate the statistical features comparing with packet sizes. They firstly extract packet sizes of first six packets and its derived statistical features on three data sets. And then using mutual information analysis they computed the effectiveness of each features and corresponding label traffic. They used equation 2 for mutual information analysis and the relationship between entropy and mutual information figure are given in Figure 2. Through mutual information analysis they develop MI features data set. Using ten well-known machine learning classifier and crossover identification experiments with features sets they achieve very high accuracy results. But the mean purpose of this research was that to find out the effectiveness of packet size and statistical features. They showed that global features, statistical features are effective features as the pay load size s for early stage traffic identification. They also found that the minimum features is not effective features or early stage traffic classification. Thus they concluded that feature should be selected carefully for early internet traffic identification.

$$I(X;Y) = \int_Y \int_X p(x,y)(\frac{p(x,y)}{p(x)p(y)})dxdy \tag{3}$$

The above given equation 3 use by the author for mutual information analysis to find out the effectiveness of features.

10. Effective Packet Number for Early Stage Internet Traffic Classification In 2015 again Peng L et al. [127] another published another paper related early stage internet traffic classification. In this study work authors focused on packet number for early traffic classification. They want to know, how many packets are effective for early stage network traffic classification. For this purpose they used three data sets and the sizes of first ten packets are extracted for this study. To find the information of first ten packets

sizes they used mutual information analysis of flow type. Moreover they used correlation analysis to find out the features redundancies and they execute a number of crossover identification experiment using 11 supervised machine learning classifiers and at the end they applied statistical test to out which number of packet perform best. And has experiment showed that 5-7 packet numbers is the best for early stage internet traffic classification.

11. Imbalanced Traffic Identification Using an Imbalanced Data Gravitation Based Classification Model Peng Let al. [3] in 2016 proposed an IDGC-based model resolve the problem of imbalance internet traffic classification. For this purpose they developed firstly six imbalanced traffic dataset from the original data sets and then they extract their early stage traffic features accord with packets sizes. And then in identification experiments they compared DGC and six standard algorithms and four imbalanced algorithms with IDGC. They showed that IDGC perform very well and also standard classification can achieve effective accuracy with imbalanced traffic data set. But they mentioned in has paper that their performance is not very well as well as generalizability problem. They also found that C4.5 Performed well in the experiments. And at the end they concluded that single model is not effective for imbalanced traffic classification.

There are many ways to identify and classify internet anomaly traffic accurately. But nowadays early traffic classification is very hot topic in the network traffic classification field. In this survey paper, we review 11 paper related to early internet traffic classification. Which are recently published in 2011-2018. In this review paper, we review different approaches for early traffic classification. We found that almost papers for early stage traffic classification are related to features evaluation such as in 2014-2018 all the papers that we review about features evaluation for early traffic classification. The summary of review papers are given in Table 6 which consist on reviews paper detail such as references, paper publication year, name of dataset used in paper, features that are

used in paper and in last method that how many methods are use and name of method used in paper. However early traffic classification is effective for internet traffic classification but their still need improvement related to features selection. It means that the feature selection technique still needs improvement for effectively early traffic classification in and more methods should be applied to select efficient feature, this is the main future work.

6. ESSENTIAL CONCEPT OF MEASUREMENT TECHNIQUE & PERFORMANCE EVALUATION CRITERIA

Internet/ Network traffic classification or identification is not easy task as well as it is also very difficult to understand the behavior of computer networks[128, 129]. However there are many techniques to classify internet traffic online and offline and also several techniques to measure the traffic. We have discussed these two major techniques in detail below and details can find in figure 6.

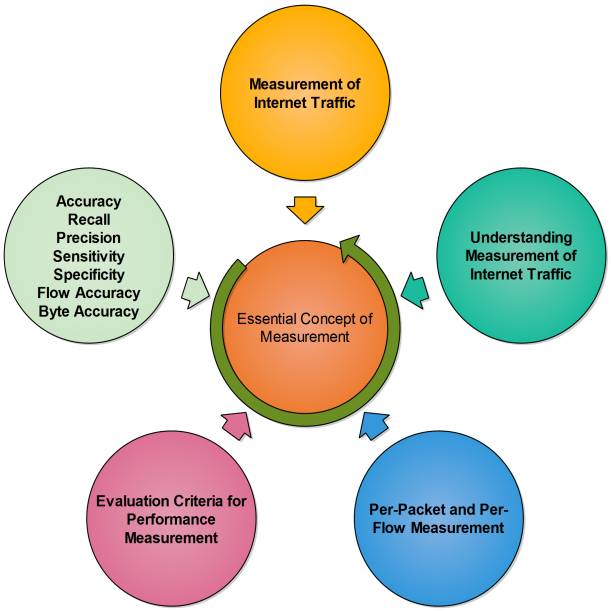

Figure 5: Understanding Measurement Metrics

6.1. Measurement of Internet Traffic

Network traffic measurement is also very crucial task in scientific research community. As MCGregor in [130] study network measurement and mentions some technical challenges in order to conducting quality measurements.

6.2. Understanding Measurement of Internet Traffic

To understand the computer network traffic behavior accurately and conduct quality measurements Williamson in [129] categorized the research tools for the purpose of network study as Online & Offline, Hardware & Software, Protocol level, and Active & Passive. The detail explanation of the above given each category are given in below.

- Online approach refers to the analyzing of traffic while the traffics are flowing in network. This process is very effective process and useful in application such as firewalls in order to block some unwanted traffic in a network. But it is need very expensive resources in high speed network.

- Offline In this approach the traffic is capture and saves for the offline analysis at later time. Its mean that analyses are conducted when the packet has been crossed the network. This technique mostly effective for researcher and when the analysis is not need online. On that data that are capture for offline analysis can be used in several approaches.

- Hardware For better performance solution dedicated hardware is also useful of the analysis of real time internet traffic. For this purpose of real time internet traffic analysis some companies provide hardware base solution such as Wildpackets [131] Napatech [132] and ipoque [133] etc.

- Protocol Level Internet traffic measurement can be performed on different level single protocol level or multiple protocol level. But mostly internet traffic is measure using IP level or Ethernet level by the researchers.

- Active This approach refers to the analysis of actual packets in the network to analysis the behavior or statistics of the traffic. This approach

controls the flows of traffic in network. But this approach is having some disadvantages such as it affect the performance of network. Moreover this does not reflect the behavior of the actual traffic flowing in the traffic. This means that it is not very well for extracting behavior of flowing traffic.

- Passive This approach is alternative to the active approach because this approach does not analysis the actual packet in the network to analyze the behavior of network traffic and this approach does not affect the performance of network mean bandwidth. But the main problem in this approach is that this approach processes a large size of data to achieve effective information of network traffic.

- LAN and MAN Mostly measurement are applied in LAN instead of WAN, because of there is need to lose the information.

- Software As we know that mostly traffic are classified of IP traffic or Ethernet frames network, so there is not matter whether traffic are classified using hardware based or software based.

6.3. Per-Packet and Per-Flow Measurement

In the field of network traffic classification mostly researcher study IP packet or we can say Ethernet frames. While Per packet approach per packet is captured flowing in computer network and then after capture the packet it is analysis for further information. This approach is useful in network security such as anomaly detection, intrusion detection System. After analyzing the packet some useful decision can be made. Moreover the packet that is capture for analysis can be used to store for offline analysis. For this purpose packet analyzer tool can be used like Wireshark [40] and Ettercap [134]. These tools are packet capturing and analyzing application. While flow means a set of packets flowing in a network and also shares common 5 tuple characteristics, where 5 tuple means Source-IP, Destination-IP, Destination Port and protocol. Where flow can unidirectional or bidirectional such as if no differentiation is made between packets travelling in each direction, we will called the flow is unidirectional flow

otherwise flow will be bidirectional. For the measurement of network traffic and bandwidth management unidirectional flows is effective and for TCP session and traffic classification purpose bidirectional is effective in which traffic are flowing in two sides. Bidirectional flow approach is also effective for video and audio traffic classification. There are many tools for flow for flow based analysis but the most important tool are Coral-Reef [135], Netflow [136] and IPFIX [**?**] etc.

6.4. Evaluation Criteria for Performance Measurement

In network traffic classification technique process at the end the results of utilized method are evaluated through specific metrics defends on the researcher which metrics he wants to use for the performance measurements. However confusion matrix is the base of traffic classification measurements. Figure 7 shows confusion matrix for traffic classification performance evaluation. Where in rows refer to the actual class of the instances and column refers to the predicted class of instances.

The metrics that are used in internet traffic classification using confusion metrics are describe below step by step.

- True Positive (TP): Its means that a Class A is truly identified as belonging to a Class A.
- True Negative (TN): Its means that a Class A is truly identified as not belonging to Class A.
- False Positive (FP): Its means that a Class A is not truly identified as belonging to Class A.
- False Negative (FN): Its means that a Class A is not truly identified as not belonging to Class A.

Using above given metrics different metrics can be made for the evaluation of classification performance [137], [138]. But note that effective classifiers will minimize the FP and FN values. The metrics that are commonly used in Internet traffic classification are given with details as shown in figure 5.

Table 4: Summary of selected papers for early traffic classification

Reference	Paper	Dataset	Features	Method
[19]	Hullr et al	They used has own captured traffic dataset	First 16 bytes of the first packet of each flow	Random Forest or Context Tree
[18]	Dainotti et al	They used has own captured traffic dataset	PS, IPT, Average Standard Deviation, L4 Protocol, Biflow duration and size, IPT statistics, Payload, Ports	J48, K-NN, R-TR, RIP, MLP, NBAY, PL, PORT
[21]	Dainotti et al	—	—	—-
[23]	Nguyen et al	They used has own captured traffic dataset	Packet length and Packet inter arrival time	Nave Bays and C4.5 Decision Tree
[22]	Qu at al	UNIBS 2009 anonym zed Internet Trace and UNIBS 2009 SSH-tunnel traces	Packet Direction Pattern, Size and Inter Arrival Time	C4.5 Decision Tree, Support Vector Machine and Nave Bayes.
[24]	Huang, N at al	They used has own captured traffic dataset	Max, Min, Median and Average statistical features.	APPR and C4.5 Decision Tree. PART, Bayes NET, Nave Bayes, OneR and ZeroR
[25]	Rizzi et al.	They used has own captured traffic dataset	Flow direction, length, and timestamp	PARC Algorithm and SVM Algorithm
[26]	Peng L et al.	Auckland II, UNIBS and UJN traces	Packet size, Average, Variance, Standard Deviation, Geometric Means Max and Minimum as well as Diff1 and Diff2	Bayes Net, KNN, Bagging, PART, C4.5, Random Forest and Logistic
[27]	Peng L et al	Auckland II, UNIBS and UJN traces	Packet size, Average, Variance, Standard Deviation, Geometric Means Max and Minimum	NaveBayes, Bayes Net, Adaboost, Bagging, OneR, PART, KNN, J48, NBTree, and RandomForest
[28]	Peng L et al	Auckland II, UNIBS and UJN traces	Packet size and its derived statistical features	NaiveBayes, AdaBoost, Logistic, OneR, RanodmForest, BayesNet, Bagging, NBTree, RBFNetwork, PART and SVM
[29]	Peng Let al	Auckland II, UNIBS and UJN traces	First six packet as a features	Adaboost, Bagging, C4.5cs, IDGG, KNN, PNN, RandomForest, MOTE-Baggin, SMOTEBoost ans SVMCS

Table 5: Summary of selected papers for anomaly and instrusion detection using ML technique

Reference	Paper	Method	Application	ACC
[123]	Detection of network attacks	K-means algorithm, Semi Supervised Algorithm	TCP/IP	80.10%
[124]	Intrusion detection system	SVM, MCLPDR	-	97.23%
[125]	Intrusion detection system	OS-ELM	TCP/IP, UDP, TCP, ICMP	98.66%
[126]	Intrusion detection system	SVM TCP		99%
[127]	Intrusion detection system	Random Forest	-	99.68%
[128]	A technique for anomaly detection system	LS-SVM	TCP/IP	99.62%
[129]	Major challenges facing IoTs	-	TCP/IP, RPL	-
[130]	Real-time hybrid intrusion detection	OPForest, Clustering, SA-IDSs	RPL, 6LoWPAn	96.02%
[131]	Anomaly detection method	SVM, Nave Bayes, J48	IP	-
[132]	Network intrusion detection technique	Random Forest, SVM	TCP, UDP, ICMP	99%
[133]	Intrusion detection System	-	TCP/IP	-
[134]	Network anomaly detection	SVM, ADAM, OCSVM	TCP/IP, UDP, ICMP	-
[135]	Security measurement	-	CoAP, HTTPS, HTTP, TLS	-
[136]	An overview of IoT	-	IPV4/IPV6	-
[137]	Brief overview of IoT	-	TCP/IP, IPv6, RPL	-

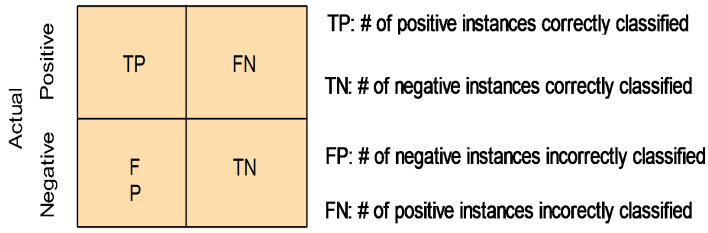

Figure 6: Confusion Matrix for Result Evaluation

- Accuracy Classification accuracy can be defined as the truly classified samples in overall classified samples. And its formula is given below. In mathematically accuracy can by define as the sum of TP and TN divide by the sum of TN, TN, FP and FN.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{4}$$

A classifier performance is measure by accuracy result. It shows the overall effectiveness of classification model.

- Recall Classification recall can be defined as the percentage sample of a particular class A classified as belonging to class A. Where the Equation 2 shows its formula. In mathematically it can be defined as TP divide by sum of TP and FN. Remember that recall metrics uses for measure the positive classes present in the dataset.

$$Recall = TP/(TP + FN) \tag{5}$$

- Precision It can be defined as the percentage of samples which are truly Class A in all those were identified class A. Equation 6 show its formula while mathematically it be defined as the TP divide by sum of TP and FP. Note the in classification technique recall and precision metrics are combine used to measure the classifier performance.

$$Precission = TP/(TP + FP) \tag{6}$$

- Sensitivity Remember that sensitivity and recall are the same metrics in traffic classification technique. So Equation 7 can be used for sensitivity.

$$Sensitivity = TP/(TP + FN) \tag{7}$$

- Specificity It can be defined as the performance ability of machine learning classifier to classify negative results. Equation 5 shows it formula and while mathematically it can be defined as TN divide by sum of FP and TN.

$$Specificity = TN/(FP + TN) \tag{8}$$

- Flow Accuracy In internet traffic classification, its means that the number of truly classified flows divided by total number of flows.
- Bytes Accuracy It can be defined as a number of correctly classified flow bytes divided by total number of flows bytes.

Flow accuracy and Bytes accuracy are mainly used in flow base internet traffic classification. Flow accuracy also mean that the measurement of accuracy: how many flows are classified in total flows. While byte accuracy is measurement of accuracy that how many bytes carried by the packets in total number of packets. Erman et al. in [139] argue that byte accuracy in network traffic classification is very difficult for evaluating accuracy of machine learning classifier. They also argue that in network traffic a majority of flows are flowing with very small bytes size and with small portion of total bytes. Hence these are the most commonly used measurement metrics used in anomaly Internet traffic classification to measure the performance of specific classifier.

The main aim of this work is to discuss the assortment of classification methods for anomaly identification. For the review paper, we choose those papers which were published recently (2011- 2018) on early traffic classification. For this study we review those research work which were newly published and in our opinion interesting. In our study knowledge there is no survey study about early stage internet traffic classification has been studied yet. This is the first

survey paper on early stage internet traffic classification. The paper that we choose for this survey paper are very interesting as well give you quick insight for extracting features for early stage network traffic classification. As well as using these tool to trace the traffic and analyzing the trace traffic are required more space and high level resources, however to overcome this problem trace reduction can be used to which reduce the size of trace traffic by applying packet filtering methods.

7. OBSERVATION AND RECOMMENDATIONS

The given literature reviews and papers found on ML and DM techniques for Sustainable Smart Cities traffic classification shows that these techniques are very effective and growing research area for Internet traffic classification, network management, Quality of service and network security etc. However, many questions arise here in machine learning technique, data set, features selection and extraction etc. Method is effective for accurate Internet work traffic classification using machine learning techniques. Unfortunately, this not yet considered.

7.1. Observation Related to the Data Sets

In this survey paper, the most cited and used data set are for anomaly detection and Internet traffic classification using machine learning algorithm for Cyber Security mean misuse detection as well as at every aspect related to internet traffic classification discus. Many researcher used the above mention dataset in their research, but making accurate and effective the dataset is very difficult job. Whenever a publically accurate data set is available, researchers like to reuse the available datasets. This is not good habit for researchers. However, the available publically datasets are good for comparison of experimental results with own developed datasets and allow you to compare your proposed technique. Thus, it is very important to develop a new data set while reusing publically available dataset. Moreover, some researches of machine learning (ML) based

traffic classification use trace traffic sample on key nodes for network for their research study. In fact these collected samples do not have accurate traffic classification / identification information, such as most important is ground truth which is very difficult for ML algorithms. Thus a model should be developed for the developing of accurate data set and to collect traffic samples with accurate application information's.

7.2. Observation Related to Feature Selection & Extraction

Classification of Internet traffic accurately is very important for the application of traffic classification or identification. And this has nowadays become very hot topic and caught a lot of researchers in recent years. Because effective feature selection provides a way to reduce the computational complexity mean to reduce the computational time. Feature selection also provides a better understanding of the data in machine learning or pattern recognition applications. However, feature selection is a technique to select subset of variable from the input. Moreover, feature selection describes the input and reduces the noise or irrelevant variables. For this purpose many researches proposed different model for feature selection and got effective results as we discussed in section V. However, it is important to find out and evaluates the effectiveness of selected features. Therefore, different feature set should be build using different feature selection techniques to find out the effectiveness of features. This is the one topic for future research study. It is very important to find out the difference and effectiveness between features for Internet traffic classification. Secondly, from the global view all the developed features of number of dataset should be compared to find out the most effective features from the global features sets. Similarly, a new feature selection technique should be developed to test the selected features selected from the global feature set.

7.3. Observation Related to ML for Intrusion Internet Traffic Classification

Machine learning (ML) algorithms are very important for many application traffic classifications in Internet traffic classification. Although, all the applied

machine learning algorithms are very useful for traffic identification as we discuss in section 5 respective. Although these ML and DM methods are useful for network traffic classification such some ML/DM algorithms are based on statistical and some are entropy like decision and some are evolutionary etc. However, it is also important to know that the training data set is useful mean effective statistical properties. Moreover, it is very important to know that the model proposed will work online or offline. After more, there still no study which showed the most effective machine learning classifier for network traffic classification. A new survey of tutorials should be proposed to find out the most effective machine learning classifier for Internet traffic classification. The study should be not limited to limited machine learning classifier neither limited to dataset. However, dataset should be selected related to only Internet traffic classification.

7.4. Observation Related to ML & DM Intrusion Detection System Performance

As we discussed that for accurate intrusion or anomaly detection effective dataset is very important and we also discussed that numerous researcher used the same data sets for their research study such as DARPA or KDD data sets and used machine learning algorithm for effective identification. However, their studies did not proposed accurate intrusion detection system or anomaly detection systems, they just analysis the performance of different machine learning algorithms on different anomaly or intrusion dataset. Therefore, a new and different from the other proposed IDSs should be proposed and then compared different machine learning algorithm performance results with other proposed IDSs. Moreover, several researcher studies and utilized the traditional machine learning algorithm for anomaly detection and they did not proposed the machine learning algorithm. For instance, a few researcher studies and utilized Reinforcement Learning, Deep RL and Game theory. Though, these are very effective learning method. But, a few researcher used these learning technique. Thus, a new ML algorithm should proposed and also should be used different learning technique such as Reinforcement and Deep Reinforcement learning

method etc.

7.5. Observation Related to Early Traffic Classification: Effective Pakcet Number

Although accurate Internet traffic classification is very important for network application traffic classification and intrusion detection traffic identification. From last few years many researchers try hard to present accurate machine learning model for Internet traffic classification and intrusion detection systems. However, different techniques and methods are proposed to for the accurate Internet traffic in which early stage Internet traffic was one effective method. Nowadays Early Stage Internet Traffic Classification got very importance in traffic identification. However, recently different techniques are proposed for early stage traffic classification. Although effective packet number for early stage network traffic classification is one of them and some researcher study and find out effective packet number for early stage network traffic classification. Some researcher study that early ten packet is enough and is very effective for traffic classification while some showed that only early six packets is good and as well as some mentioned first 19 packet is very effective for early stage Internet traffic classification. However, this problem still to be studied to find out the exact and effective packet number for early stage Internet traffic classification and should be compared with the previous published authors works. Similarly feature evaluation and selection also should be discuss and study, because it is also need to study in depth. However, a new method should be proposed to find effective feature set for early stage Internet traffic classification.

8. Conclusion

In this paper, we describe the literature review of sustainable smart cities, machine learning (ML), data mining (DM), anomaly and intrusion detection systems used for Internet traffic classification. Importance was placed with example papers that describe the use of ML and DM method for sustainable

smart cities network traffic classification. Different data sets were discuses with feature selection and extraction. However, we select those ML, DM, datasets and discuss with details that are mostly used by researchers for research study in Internet anomaly traffic classification. For better understanding, a short tutorial of Internet traffic classification using ML/DM methods with a selected dataset of the feature is provided. Considering relevance and most cited methods and datasets of features were identified, read and summarized. As data and data features are essential in Internet traffic classification using ML/DM technique, some well-known and most used dataset with details statistical features are also described. Different classification techniques for network anomaly traffic classification are presented with details. The complexity of data set, features extraction and ML/DM methods are addressed. At the end challenges and recommendations for Internet traffic classification using ML/DM technique with the dataset of features are presented. Similarly, a recommendation also presents for the future, but it is very crucial to make one recommendation for each method. However, some recommendation related to traffic classification are presented in section VIII which is very significant for future work.

References

[1] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, C. Wang, Machine learning and deep learning methods for cybersecurity, IEEE Access 6 (2018) 35365–35381.

[2] A. Saeed, M. Kolberg, Towards optimizing wlans power saving: Novel context-aware network traffic classification based on a machine learning approach, IEEE Access 7 (2019) 3122–3135.

[3] L. Peng, H. Zhang, Y. Chen, B. Yang, Imbalanced traffic identification using an imbalanced data gravitation-based classification model, Computer Communications 102 (2017) 177–189.

[4] H. N. Datir, P. M. Jawandhiya, Survey on hybrid data mining algorithms

for intrusion detection system, in: Data Management, Analytics and Innovation, Springer, 2019, pp. 291–298.

[5] M. Shafiq, Z. Tian, Y. Sun, X. Du, M. Guizani, Selection of effective machine learning algorithm and bot-iot attacks traffic identification for internet of things in smart city, Future Generation Computer Systems.

[6] K. Hayat, M. I. Ali, J. C. R. Alcantud, B.-Y. Cao, K. U. Tariq, Best concept selection in design process: An application of generalized intuitionistic fuzzy soft sets, Journal of Intelligent & Fuzzy Systems 35 (5) (2018) 5707–5720.

[7] T. Karagiannis, A. Broido, M. Faloutsos, et al., Transport layer identification of p2p traffic, in: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, ACM, 2004, pp. 121–134. doi: 10.1109/MASCOTS.2006.6.

[8] S. Dong, R. Jain, Flow online identification method for the encrypted skype, Journal of Network and Computer Applications 132 (2019) 75–85.

[9] S. Sen, O. Spatscheck, D. Wang, Accurate, scalable in-network identification of p2p traffic using application signatures, in: Proceedings of the 13th international conference on World Wide Web, ACM, 2004, pp. 512–521.

[10] G. Ren, Y. Cao, S. Wen, T. Huang, Z. Zeng, A modified elman neural network with a new learning rate scheme, Neurocomputing 286 (2018) 11–18.

[11] S. Aljawarneh, M. Aldwairi, M. B. Yassein, Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model, Journal of Computational Science 25 (2018) 152–160.

[12] M. Piskozub, R. Spolaor, I. Martinovic, Malalert: Detecting malware in large-scale network traffic using statistical features, ACM SIGMETRICS Performance Evaluation Review 46 (3) (2019) 151–154.

[13] L. Wang, X. Wang, M. Tornatore, K. J. Kim, S. M. Kim, D.-U. Kim, K.-E. Han, B. Mukherjee, Scheduling with machine-learning-based flow detection for packet-switched optical data center networks, Journal of Optical Communications and Networking 10 (4) (2018) 365–375.

[14] R. Chapaneri, S. Shah, A comprehensive survey of machine learning-based network intrusion detection, in: Smart Intelligent Computing and Applications, Springer, 2019, pp. 345–356.

[15] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, E. Vázquez, Anomaly-based network intrusion detection: Techniques, systems and challenges, computers & security 28 (1-2) (2009) 18–28.

[16] K. A. Taher, B. M. Y. Jisan, M. M. Rahman, Network intrusion detection using supervised machine learning technique with feature selection, in: 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), IEEE, 2019, pp. 643–646.

[17] T. T. Nguyen, G. J. Armitage, A survey of techniques for internet traffic classification using machine learning., IEEE Communications Surveys and Tutorials 10 (1-4) (2008) 56–76.

[18] A. L. Buczak, E. Guven, A survey of data mining and machine learning methods for cyber security intrusion detection, IEEE Communications Surveys & Tutorials 18 (2) (2016) 1153–1176.

[19] A. C. Callado, C. A. Kamienski, G. Szabó, B. P. Gero, J. Kelner, S. F. Fernandes, D. F. H. Sadok, A survey on internet traffic identification., IEEE Communications Surveys and Tutorials 11 (3) (2009) 37–52.

[20] M. Bhatia, M. K. Rai, Identifying p2p traffic: A survey, Peer-to-Peer Networking and Applications 10 (5) (2017) 1182–1203.

[21] J. Hurley, E. Garcia-Palacios, S. Sezer, Classification of p2p and http using specific protocol characteristics, in: Meeting of the European Net-

work of Universities and Companies in Information and Communication Engineering, Springer, 2009, pp. 31–40.

[22] F. López Rodríguez, U. Silva Dias, D. R. Campelo, R. d. Oliveira Albuquerque, S.-J. Lim, L. J. García Villalba, Qos management and flexible traffic detection architecture for 5g mobile networks, Sensors 19 (6) (2019) 1335.

[23] M. Mohammadi, B. Raahemi, A. Akbari, H. Moeinzadeh, B. Nasersharif, Genetic-based minimum classification error mapping for accurate identifying peer-to-peer applications in the internet traffic, Expert Systems with applications 38 (6) (2011) 6417–6423.

[24] M. Narayanan, An efficient method to classify the peer-to-peer network videos and video servers over video on demand services, in: Innovations in Electronics and Communication Engineering, Springer, 2019, pp. 187–195.

[25] H. Schulze, K. Mochalski, Internet study 2008/2009, Ipoque Report 37 (2009) 351–362.

[26] H. Abbasi, N. Ezzati-Jivan, M. Bellaiche, C. Talhi, M. R. Dagenais, Machine learning-based edos attack detection technique using execution trace analysis, Journal of Hardware and Systems Security (2019) 1–13.

[27] K. Zeilenga, Internet assigned numbers authority (iana) considerations for the lightweight directory access protocol (ldap), Tech. rep. (2002).

[28] M. Roughan, S. Sen, O. Spatscheck, N. 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, ACM, 2004, pp. 135–148.

[29] M. H. Mazhar, Z. Shafiq, Real-time video quality of experience monitoring for https and quic, in: IEEE INFOCOM 2018-IEEE Conference on Computer Communications, IEEE, 2018, pp. 1331–1339.

[30] A. W. Moore, K. Papagiannaki, Toward the accurate identification of network applications, in: International Workshop on Passive and Active Network Measurement, Springer, 2005, pp. 41–54.

[31] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an, H. Ye, Significant permission identification for machine-learning-based android malware detection, IEEE Transactions on Industrial Informatics 14 (7) (2018) 3216–3225.

[32] V. Paxson, Empirically-derived analytic models of wide-area tcp connections.

[33] A. K. Marnerides, D. P. Pezaros, D. Hutchison, Internet traffic characterisation: Third-order statistics & higher-order spectra for precise traffic modelling, Computer Networks 134 (2018) 183–201.

[34] T. Lang, G. Armitage, P. Branch, H.-Y. Choo, A synthetic traffic model for half-life, in: Australian Telecommunications Networks & Applications Conference, Vol. 2003, 2003.

[35] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A. E. Samir, O. S. Pianykh, J. R. Geis, P. V. Pandharipande, J. A. Brink, K. J. Dreyer, Current applications and future impact of machine learning in radiology, Radiology 288 (2) (2018) 318–328.

[36] T. Lang, P. Branch, G. Armitage, A synthetic traffic model for quake3, in: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology, ACM, 2004, pp. 233–238.

[37] W. Niu, Z. Zhuo, X. Zhang, X. Du, G. Yang, M. Guizani, A heuristic statistical testing based approach for encrypted network traffic identification, IEEE Transactions on Vehicular Technology.

[38] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The kdd process for extracting useful knowledge from volumes of data, Communications of the ACM 39 (11) (1996) 27–34.

[39] C. Shearer, The crisp-dm model: the new blueprint for data mining, Journal of data warehousing 5 (4) (2000) 13–22.

[40] A. Orebaugh, G. Ramirez, J. Beale, Wireshark & Ethereal network protocol analyzer toolkit, Elsevier, 2019.

[41] V. Jacobson, C. Leres, S. McCanne, Tcpdump manual page. lawrence berkeley national laboratory, University of California, Berkeley, CA.

[42] G. F. Lyon, Nmap network scanning: The official Nmap project guide to network discovery and security scanning, Insecure, 2009.

[43] A. Dupuy, S. Sengupta, O. Wolfson, Y. Yemini, Design of the netmate network management system, in: Integrated Network Management, Vol. 2, Elsevier Science-North Holland, 1991.

[44] L. F. Shampine, M. W. Reichelt, The matlab ode suite, SIAM journal on scientific computing 18 (1) (2019) 1–22.

[45] S. R. Garner, et al., Weka: The waikato environment for knowledge analysis, in: Proceedings of the New Zealand computer science research students conference, 1995, pp. 57–64.

[46] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, F. Abdessamia, Network traffic classification techniques and comparative analysis using machine learning algorithms, in: 2016 2nd IEEE International Conference on Computer and Communications (ICCC), IEEE, 2016, pp. 2451–2455.

[47] L. Peng, H. Zhang, B. Yang, Y. Chen, M. T. Qassrawi, G. Lu, Traffic identification using flexible neural trees, in: 2010 IEEE 18th International Workshop on Quality of Service (IWQoS), IEEE, 2010, pp. 1–5.

[48] W. Han, J. Xue, H. Yan, Detecting anomalous traffic in the controlled network based on cross entropy and support vector machine, IET Information Security 13 (2) (2019) 109–116.

[49] M. Alllman, V. Paxson, Issues and etiquette concerning use of shared measurement data, in: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, 2007, pp. 135–140.

[50] V. Jacobson, C. Leres, S. McCanne, Tcpdump/libpcap, Accessed: Jun 23 (1987) 2016.

[51] J. Frankel, D. Greely, B. Sawyer, MP3 Power! With Winamp, Thomson Course Technology, 1999.

[52] X. Huang, J. Wang, S. Aluru, S.-P. Yang, L. Hillier, Pcap: a whole-genome assembly program, Genome research 13 (9) (2003) 2164–2170.

[53] W. N. R. Group, et al., Waikato internet traffic storage.

[54] M. Bennasar, Y. Hicks, R. Setchi, Feature selection using joint mutual information maximisation, Expert Systems with Applications 42 (22) (2015) 8520–8532.

[55] A. Shiravi, H. Shiravi, M. Tavallaee, A. A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, computers & security 31 (3) (2012) 357–374.

[56] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, A detailed analysis of the kdd cup 99 data set, in: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, IEEE, 2009, pp. 1–6.

[57] W. N. R. Group, et al., Wits: Waikato internet traffic storage, URL: http://wand. net. nz/wits/index. php.

[58] L. Peng, B. Yang, Y. Chen, Z. Chen, Effectiveness of statistical features for early stage internet traffic identification, International Journal of Parallel Programming 44 (1) (2016) 181–197.

[59] M. H. Bhuyan, D. K. Bhattacharyya, J. K. Kalita, Towards generating real-life datasets for network intrusion detection., IJ Network Security 17 (6) (2015) 683–701.

[60] F. Gringoli, L. Salgarelli, M. Dusi, N. Cascarano, F. Risso, et al., Gt: picking up the truth from the ground for internet traffic, ACM SIGCOMM Computer Communication Review 39 (5) (2009) 12–18.

[61] Y.-s. Lim, H.-c. Kim, J. Jeong, C.-k. Kim, T. T. Kwon, Y. Choi, Internet traffic classification demystified: on the sources of the discriminative power, in: Proceedings of the 6th International COnference, ACM, 2010, p. 9.

[62] L. Chaddad, A. Chehab, I. H. Elhajj, A. Kayssi, Mobile traffic anonymization through probabilistic distribution, in: 2019 22nd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), IEEE, 2019, pp. 242–248.

[63] M. Wasikowski, X.-w. Chen, Combating the small sample class imbalance problem using feature selection, IEEE Transactions on knowledge and data engineering 22 (10) (2010) 1388–1400.

[64] A. Moore, D. Zuev, M. Crogan, Discriminators for use in flow-based classification, Tech. rep. (2013).

[65] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, X. Yu, Feature selection for optimizing traffic classification, Computer Communications 35 (12) (2012) 1457–1471.

[66] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, K. Salamatian, Traffic classification on the fly, ACM SIGCOMM Computer Communication Review 36 (2) (2006) 23–26.

[67] S. Dey, Q. Ye, S. Sampalli, A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks, Information Fusion 49 (2019) 205–215.

[68] S. Ding, H. Zhu, W. Jia, C. Su, A survey on feature extraction for pattern recognition, Artificial Intelligence Review 37 (3) (2012) 169–180.

[69] Z. Chen, L. Peng, S. Zhao, L. Zhang, S. Jing, Feature selection toward optimizing internet traffic behavior identification, in: International Conference on Algorithms and Architectures for Parallel Processing, Springer, 2014, pp. 631–644.

[70] W. Zhou, L. Dong, L. Bic, M. Zhou, L. Chen, Internet traffic classification using feed-forward neural network, in: 2011 International Conference on Computational Problem-Solving (ICCP), IEEE, 2011, pp. 641–646.

[71] R. Sun, B. Yang, L. Peng, Z. Chen, L. Zhang, S. Jing, Traffic classification using probabilistic neural networks, in: 2010 Sixth International Conference on Natural Computation, Vol. 4, IEEE, 2010, pp. 1914–1919.

[72] P. Flach, Performance evaluation in machine learning: The good, the bad, the ugly and the way forward.

[73] C. Witt, M. Bux, W. Gusew, U. Leser, Predictive performance modeling for distributed batch processing using black box monitoring and machine learning, Information Systems.

[74] A. Bivens, C. Palagiri, R. Smith, B. Szymanski, M. Embrechts, et al., Network-based intrusion detection using neural networks, Intelligent Engineering Systems through Artificial Neural Networks 12 (1) (2002) 579–584.

[75] M. Cui, J. Wang, M. Yue, Machine learning based anomaly detection for load forecasting under cyberattacks, IEEE Transactions on Smart Grid.

[76] D. Heckerman, A tutorial on learning with bayesian networks, in: Innovations in Bayesian networks, Springer, 2008, pp. 33–82.

[77] D. Heckerman, A tutorial on learning with bayesian networks, in: Innovations in Bayesian networks, Springer, 2008, pp. 33–82.

[78] C. Livadas, R. Walsh, D. E. Lapsley, W. T. Strayer, Using machine learning techniques to identify botnet traffic., in: LCN, Citeseer, 2006, pp. 967–974.

[79] S. Maeda, A. Kanai, S. Tanimoto, T. Hatashima, K. Ohkubo, A botnet detection method on sdn using deep learning, in: 2019 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2019, pp. 1–6.

[80] T. Auld, A. W. Moore, S. F. Gull, Bayesian neural networks for internet traffic classification, IEEE Transactions on neural networks 18 (1) (2007) 223–239.

[81] C.-L. Hsieh, N. Weng, W. Wei, Scalable many-field packet classification for traffic steering in sdn switches, IEEE Transactions on Network and Service Management 16 (1) (2019) 348–361.

[82] C. Kruegel, D. Mutz, W. Robertson, F. Valeur, Bayesian event classification for intrusion detection, in: 19th Annual Computer Security Applications Conference, 2003. Proceedings., IEEE, 2003, pp. 14–23.

[83] M. Wu, Z. Song, Y. B. Moon, Detecting cyber-physical attacks in cyber-manufacturing systems with machine learning methods, Journal of intelligent manufacturing 30 (3) (2019) 1111–1123.

[84] J. Gao, S. Chai, B. Zhang, Y. Xia, Research on network intrusion detection based on incremental extreme learning machine and adaptive principal component analysis, Energies 12 (7) (2019) 1223.

[85] J. Quinlan, Induction of decision trees. mach. learn.

[86] S. L. Salzberg, C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, Machine Learning 16 (3) (1994) 235–240.

[87] M. Shafiq, X. Yu, Effective packet number for 5g im wechat application at early stage traffic classification, Mobile Information Systems 2017.

[88] M. Shafiq, X. Yu, A. A. Laghari, D. Wang, Effective feature selection for 5g im applications traffic classification, Mobile Information Systems 2017.

[89] Y. Freund, R. E. Schapire, et al., Experiments with a new boosting algorithm, in: icml, Vol. 96, Citeseer, 1996, pp. 148–156.

[90] C. Xu, H. Chen, J. Wang, Y. Guo, Y. Yuan, Improving prediction performance for indoor temperature in public buildings based on a novel deep learning method, Building and Environment 148 (2019) 128–135.

[91] J. Zhang, M. Zulkernine, A. Haque, Random-forests-based network intrusion detection systems, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38 (5) (2008) 649–659.

[92] L. Breiman, A. Cutler, Random forestsclassification description: Random forests (2014).

[93] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2016.

[94] M. Shafiq, X. Yu, A. A. Laghari, Wechat text messages service flow traffic classification using machine learning technique, in: 2016 6th International Conference on IT Convergence and Security (ICITCS), IEEE, 2016, pp. 1–5.

[95] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, F. Abdesssamia, et al., Wechat text and picture messages service flow traffic classification using machine learning technique, in: 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), IEEE, 2016, pp. 58–62.

[96] S. Owais, V. Snasel, P. Kromer, A. Abraham, Survey: using genetic algorithm approach in intrusion detection systems techniques, in: 2008 7th Computer Information Systems and Industrial Management Applications, IEEE, 2008, pp. 300–307.

[97] S. Mirjalili, J. S. Dong, A. S. Sadiq, H. Faris, Genetic algorithm: Theory, literature review, and application in image reconstruction, in: Nature-Inspired Optimizers, Springer, 2019, pp. 69–85.

[98] P. Mishra, V. Varadharajan, U. Tupakula, E. S. Pilli, A detailed investigation and analysis of using machine learning techniques for intrusion detection, IEEE Communications Surveys & Tutorials 21 (1) (2018) 686–728.

[99] X. Deng, P. Jiang, X. Peng, C. Mi, An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in internet of things, IEEE Transactions on Industrial Electronics 66 (6) (2019) 4672–4683.

[100] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu, C. Wang, Y. Liu, A survey of machine learning techniques applied to software defined networking (sdn): Research issues and challenges, IEEE Communications Surveys & Tutorials 21 (1) (2018) 393–430.

[101] P. K. Jain, R. Pamula, Two-step anomaly detection approach using clustering algorithm, in: International Conference on Advanced Computing Networking and Informatics, Springer, 2019, pp. 513–520.

[102] N. Kant, M. Mahajan, Time-series outlier detection using enhanced k-means in combination with pso algorithm, in: Engineering Vibration, Communication and Information Processing, Springer, 2019, pp. 363–373.

[103] K. Malialis, D. Kudenko, Distributed response to network intrusions using multiagent reinforcement learning, Engineering Applications of Artificial Intelligence 41 (2015) 270–284.

[104] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau, et al., An introduction to deep reinforcement learning, Foundations and Trends® in Machine Learning 11 (3-4) (2018) 219–354.

[105] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529.

[106] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, nature 529 (7587) (2016) 484.

[107] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al., Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, arXiv preprint arXiv:1806.10293.

[108] J. Gauci, E. Conti, Y. Liang, K. Virochsiri, Y. He, Z. Kaden, V. Narayanan, X. Ye, Horizon: Facebook's open source applied reinforcement learning platform, arXiv preprint arXiv:1811.00260.

[109] M. Gyanchandani, J. Rana, R. Yadav, Taxonomy of anomaly based intrusion detection system: a review, International Journal of Scientific and Research Publications 2 (12) (2012) 1–13.

[110] C. Wagner, J. François, T. Engel, et al., Machine learning approach for ip-flow record anomaly detection, in: International Conference on Research in Networking, Springer, 2011, pp. 28–39.

[111] A.-m. Yang, S.-y. Jiang, H. Deng, A p2p network traffic classification method using svm, in: 2008 The 9th International Conference for Young Computer Scientists, IEEE, 2008, pp. 398–403.

[112] A. Este, F. Gringoli, L. Salgarelli, Support vector machines for tcp traffic classification, Computer Networks 53 (14) (2009) 2476–2490.

[113] W. Li, A. W. Moore, A machine learning approach for efficient traffic classification, in: 2007 15th International Symposium on Modeling, Anal-

ysis, and Simulation of Computer and Telecommunication Systems, IEEE, 2007, pp. 310–317.

[114] A. W. Moore, D. Zuev, Internet traffic classification using bayesian analysis techniques, in: ACM SIGMETRICS Performance Evaluation Review, Vol. 33, ACM, 2005, pp. 50–60.

[115] A. Dainotti, A. Pescape, K. C. Claffy, Issues and future directions in traffic classification, IEEE network 26 (1) (2012) 35–40.

[116] A. Este, F. Gringoli, L. Salgarelli, On the stability of the information carried by traffic flow features at the packet level, ACM SIGCOMM Computer Communication Review 39 (3) (2009) 13–18.

[117] B. Hullár, S. Laki, A. Gyorgy, Early identification of peer-to-peer traffic, in: 2011 IEEE International Conference on Communications (ICC), IEEE, 2011, pp. 1–6.

[118] J. Ma, K. Levchenko, C. Kreibich, S. Savage, G. M. Voelker, Unexpected means of protocol inference, in: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, ACM, 2006, pp. 313–326.

[119] J. Khalife, A. Hajjar, J. Díaz-Verdejo, Performance of opendpi in identifying sampled network traffic, Journal of Networks 8 (1) (2013) 71.

[120] A. Finamore, M. Mellia, M. Meo, M. M. Munafo, P. Di Torino, D. Rossi, Experiences of internet traffic monitoring with tstat, IEEE Network 25 (3) (2011) 8–14.

[121] A. Dainotti, A. Pescapé, C. Sansone, Early classification of network traffic through multi-classification, in: International Workshop on Traffic Monitoring and Analysis, Springer, 2011, pp. 122–135.

[122] T. T. Nguyen, G. Armitage, P. Branch, S. Zander, Timely and continuous machine-learning-based classification for interactive ip traffic, IEEE/ACM Transactions On Networking 20 (6) (2012) 1880–1894.

[123] B. Qu, Z. Zhang, L. Guo, D. Meng, On accuracy of early traffic classification, in: 2012 IEEE Seventh International Conference on Networking, Architecture, and Storage, IEEE, 2012, pp. 348–354.

[124] N.-F. Huang, G.-Y. Jai, H.-C. Chao, Y.-J. Tzang, H.-Y. Chang, Application traffic classification at the early stage by characterizing application rounds, Information Sciences 232 (2013) 130–142.

[125] A. Rizzi, S. Colabrese, A. Baiocchi, Low complexity, high performance neuro-fuzzy system for internet traffic flows early classification, in: 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC), IEEE, 2013, pp. 77–82.

[126] L. Peng, H. Zhang, B. Yang, Y. Chen, Feature evaluation for early stage internet traffic identification, in: International Conference on Algorithms and Architectures for Parallel Processing, Springer, 2014, pp. 511–525.

[127] L. Peng, B. Yang, Y. Chen, Effective packet number for early stage internet traffic identification, Neurocomputing 156 (2015) 252–267.

[128] V. Paxson, Strategies for sound internet measurement, in: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, ACM, 2004, pp. 263–271.

[129] C. Williamson, Internet traffic measurement, IEEE internet computing 5 (6) (2001) 70–74.

[130] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, I. Rafols, Bibliometrics: the leiden manifesto for research metrics, Nature News 520 (7548) (2015) 429.

[131] M. L. Mueller, H. Asghari, Deep packet inspection and bandwidth management: Battles over bittorrent in canada and the united states, Telecommunications Policy 36 (6) (2012) 462–475.

[132] T. Kaletha, Simple wild-packets, Journal of the Institute of Mathematics of Jussieu 12 (1) (2013) 43–75.

[133] I. real-time network Analysis, Napatech.

[134] ETTERCAP, ETTERCAP (2017).

[135] D. Moore, K. Keys, R. Koga, E. Lagache, et al., Coralreef software suite as a tool for system and network administrators, in: Usenix LISA, 2001.

[136] M. Fullmer, S. Romig, The osu flowtools package and cisco netflow logs, in: Proceedings of the 2000 USENIX LISA Conference, 2000.

[137] B. Claise, B. Trammell, P. Aitken, Specification of the ip flow information export (ipfix) protocol for the exchange of flow information, Tech. rep. (2013).

[138] D. L. Olson, D. Delen, Advanced data mining techniques, Springer Science & Business Media, 2008.

[139] J. Erman, A. Mahanti, M. Arlitt, Byte me: a case for byte accuracy in traffic classification, in: Proceedings of the 3rd annual ACM workshop on Mining network data, ACM, 2007, pp. 35–38.