IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 23, 2022, accepted June 1, 2022, date of publication June 8, 2022, date of current version June 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3181135

Procedures, Criteria, and Machine Learning Techniques for Network Traffic Classification: A Survey

MUHAMMAD SAMEER SHEIKH AND YINQIAO PENG
School of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China

Corresponding author: Yinqiao Peng (pengyq@gdou.edu.cn)

This work was supported in part by the Program for Scientific Research Start-Up Funds of Guangdong Ocean University under Grant E15046, in part by the Special Project on the Key Areas of New Generation of Information Technology under Grant 2020ZDZS3008, and in part by the Special Project of Artificial Intelligence under Grant 2019KZDZS1046.

ABSTRACT Traffic classification is considered an important research area due to the increasing demand in network users. It not only effectively improve the network service identifications and security issues of the traffic network, but also provide robust accuracy and efficiency in different Internet application behaviors and patterns. Several traffic classification techniques have been proposed and applied successfully in recent years. However, the existing literature lack of comprehensive survey which could provide an overview and analysis towards the recent developments in network traffic classification. To this end, this survey presents a comprehensive investigation on traffic classification techniques by carefully reviewing existing methods from a new perspective. We comprehensively discuss the procedures and datasets for traffic classification. Additionally, traffic criteria are proposed, which could be beneficial to assess the effectiveness of the developed classification algorithm. Then, the traffic classification techniques are discussed in detail. Then, we thoroughly discussed the machine learning (ML) methods for traffic classification. For researcher's convenience, we present the traffic obfuscation techniques, which could be helpful for designing a better classifier. Finally, key findings and open research challenges for network traffic classification are identified along with recommendations for future research directions. In sum, this survey fills the gap of existing surveys and summarizes the latest research developments in traffic classification.

INDEX TERMS Classification criteria, machine learning method, obfuscation, security, traffic classification.

I. INTRODUCTION

Traffic classification is the first step that helps identify different applications and protocols that exist in the network. Several operations such as monitoring and optimization can be performed on the identified traffic with the aim to improve the network performance [1]. To identify and classify unknown network categories, traffic classification becomes an important approach due to its ability to solve various network problems and provide various solutions to Internet service providers and installed equipment [2], [3]. The network service providers are responsible for network issues and the attacks which networks are vulnerable with such as malicious activity of the user's nodes and network attacks. Traffic classification can be part of the intrusion detection system (IDS) [4], [5], which is used to detect various attacks, pattern indication of denial of service (DoS), and deployment and relocation of available network resources. Also, it aims to analyze and distinguish different types of applications flowing in the network.

In the recent years, research community focuses on ML techniques to identify traffic classification. As Peng *et al.* studied the imbalanced traffic identification and classification technique by using an imbalance data gravitation classification (IDGC) model. They constructed the numerous amount of imbalanced traffic datasets using real-world traffic datasets, and then extracted features from them by considering their actual packet size [6]. Moreover, Saeed and Kolberg studied a computational method, which helps network operators for identifying different applications [7]. Erman *et al.* proposed a semi supervised traffic classification method which is the combination of supervised and

The associate editor coordinating the review of this manuscript and approving it for publication was Shunfeng Cheng.

unsupervised ML-based techniques. The presented method overcomes the difficulty of labeling and obtaining scarce, and new application takes a longer time to detect various kind of flows [8]. As Roughan *et al.* presented a signal based method for the IP traffic classification. They used various techniques such as nearest neighboring technique, linear discriminate analysis and quadratic discriminate analysis to map the applications and determine the quality of services (QoS) of traffic classes [9]. Datir and Jawandhiya provided the essential parameters for obtaining sustainable smart cities, which are flowing throughout the network. They discussed various hybrid traffic classification approaches which are used in the intrusion detection system. These hybrid classifiers maybe in the form of any classifiers as Naive Bayes, SVM, K-means clustering, etc. [10], and indicated that various applications uses random type of port numbers in order to prevent them from the network attack or other malicious activity [11]. Ren *et al.* presented an elman neural network based on learning rate framework. They aimed to solve prediction issue at discrete time sequence [12].

For application based patterns without using packet load inspections, several network traffic classification works have been presented. It enhances the integration of different objects such as nodes and sensors, which are required to engage IoT network traffic which is different from other networks [13]. Similarly, Finsterbusch *et al.* revealed that the network traffic classification becomes the famous topic of the Internet at all stages [14]. Several new traffic classification schemes have been presented for addressing different characteristics and features such as packet arrival time and packet length, etc. [15].

Traffic classification receives significant attention from the research community. Several surveys have been conducted to address and review existing challenges, solutions and applications of traffic classification and identification. Different reviews focuses on different classification methods such as Nguyen and Armitage [3] presented a survey on traffic classification using ML-based technique. They discussed the various ML methods and IP traffic classifications and reviewed different ML methods from 2004 to 2007. It also reviewed and discussed the requirements for various ML-based classifiers. Chapaneri and Shah [16] reviewed various intrusion detection based ML techniques up to 2018. They also discussed the issues related to traditional and network intrusion datasets and outlined the open research challenges and future research directions of ML-based intrusion detection systems. Similarly, García-Teodoro *et al.* [17] mainly reviewed most popular anomaly intrusion detection techniques. They also outlined main research challenges and procedures for deploying the intrusion detection system. Callado *et al.* [18] reviewed the main techniques and issues of IP traffic and analyzed the traffic in terms of packet and flow-based categories and then outline their advantages. It also discussed the sampling and matching mechanism of signature and outlined the open research challenges of traffic analysis and application detection. Bhatia and Rai [19] launched a survey on peer-to-peer (P2P) traffic, in which they discussed the different strategies to determine P2P traffic. They also presented the analysis of traffic network measurement and monitoring.

Moreover, Buczak and Guven [20] summarized the research works related to the cyber security of ML and data mining (DM) techniques up to 2016 and outlined the open research challenges for deploying ML and DM techniques for cyber security applications. Dainotti *et al.* [21] reviewed the recent works on the traffic classification. They also discussed the various challenges that were faced by the researchers over ten years and recommended some strategies to overcome these challenges and improve the performance of the traffic classification method. Alsheikh *et al.* [22] presented a survey on the ML based technique for wireless network. Firstly, they presented a literature review on the ML-based techniques in wireless and other networks. Then, they addressed the merits and demerits of each algorithm and outlined the open research challenges for employing ML based techniques in wireless networks. Shafiq *et al.* [2] summarized the recent traffic classification methods for sustainable smart cities. They also outlined the open research challenges and proposed recommendations for traffic classification by considering the dataset features. Velan *et al.* [23] mainly reviewed existing techniques for traffic classification and analyzed the encryption protocols through the Internet. They also discussed a payload approach and feature based classification technique based on reviewing taxonomy. Gomes *et al.* [24] reviewed the peer-to-peer mechanism of traffic classification and detection techniques. They also discussed the detailed network analysis of traffic monitoring schemes. Pacheco *et al.* [25] summarized the steps to obtain the traffic classification using ML schemes. They also discussed the open research challenges and future research directions and summarized the research aim to improve the QoS and the operator network. Tahaei *et al.* [26] presented a survey on the traffic classification in the IoT network. They discussed the deployment of IoT traffic classification in real-world applications and the open research challenges in this domain.

The traffic classification and identification play significant roles to develop a better sustainable smart cities by deploying a better network management system and improving network security of the whole network. [27]. In this paper, we carry out a comprehensive review of published papers that provides various solutions for traffic classifications. The purpose of this survey is to elucidate the roadmap for those who want to do research in the traffic classification area. This survey not only discusses ML methods for traffic classification but it also discusses the traffic classification procedures and performance criteria. In particular, this survey focuses on the traffic classification techniques and the ML methods for Internet traffic classification. We classify classification techniques into four categories such as port based classification, payload based classification, statistical based classification, and behavior based classification. We discuss the

datasets for traffic classification of network-based anomaly detection in detail. These datasets could be used to evaluate the efficiency of the developed algorithms before applying them in real applications. We present the traffic classification criteria to evaluate the effectiveness of existing classification algorithms. For researcher's convenience, we present the traffic obfuscation techniques which could help them to design and develop a robust classifier, and to protect the user privacy. In the end, we outline key findings, open research challenges, and recommendations for future research directions on traffic classification. By comparing with previous surveys, we summarize the contribution of this paper as follows:

- We discuss the comprehensive literature review on the recent state-of-the-art of traffic classification methods. This literature provides useful information to the researchers and practitioners who intend to apply traffic classification in the application context.
- We comprehensively discuss the process of traffic classification, which consists of traffic datasets, features selection and extraction, and the decision and validation process. The datasets for Internet traffic classification are presented. These datasets could be helpful to assess the effectiveness of developed algorithms before applying them practically.
- We present the traffic classification criteria which can be used to assess the effectiveness of classification algorithms. It consists of effectiveness and performance criteria's. The existing traffic classification techniques are also discussed in detail.
- We comprehensively discuss the ML methods for traffic classification. We summarize the traffic classification methods and its features and applications for the convenience of other researchers and practitioners.
- We thoroughly discuss the traffic obfuscation techniques, which could be helpful for designing a better classifier.
- We discuss the key findings and various open challenges and identify issues for future research directions. These challenges reveal some useful insights that help researchers to tackle issues when employing traffic classification algorithms.

The rest of this survey is organized as follows. Section 2 discusses the procedures for traffic classification, which consists datasets for traffic classification, and extraction and selection features. Section 3 presents the criteria for traffic classification. Section 4 introduces various traffic classification techniques. Section 5 presents the ML techniques for traffic classifications. Section 6 presents traffic classification obfuscation techniques. Section 7 presents the key findings, limitations, and recommendations for employing traffic classification. Finally, Section 8 concludes the study.

II. PROCEDURES FOR TRAFFIC CLASSIFICATION

This section discusses the traffic classification process as illustrated in Figure 1. Traditional traffic network could be used as an input to establish a dataset for feature selection processing. Next, the feature extraction and selection plays a key role for traffic classification due to its effectiveness on the performance of the traffic classification. Third, the decision process (DP) could identify the class of traffic classification using the ML techniques. Finally, the validation process (VP) is used to verify the results of traffic classification by determine the accuracy of classification.

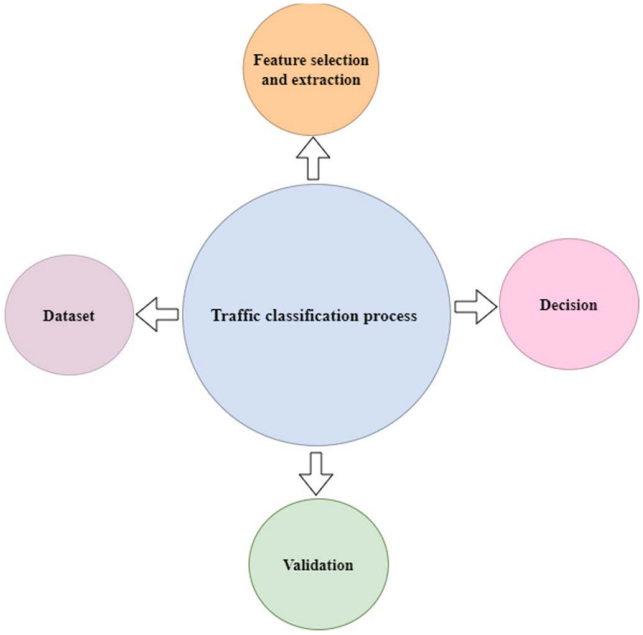

FIGURE 1. Traffic classification process.

A. DATASETS FOR TRAFFIC CLASSIFICATION

Datasets play a crucial role to assess the effectiveness and reliability of a developed algorithm. For instance, the effectiveness of Vehicular-ad-hoc-Network (VANET) and intrusion detection system (IDS) could be assessed by detecting attacks inside and outside of the network. Therefore, it requires complete datasets that consists of normal and abnormal behaviors. As the behavior and patterns of the network changes rapidly, a reliable dataset could provide an efficient mechanism to detect the traffic classification model in a real scenario.

Various numbers of datasets are available to test and evaluate different algorithms in the cybersecurity research domain. In [28], Bhuyan *et al.* discussed various datasets for cybersecurity research which are further categorized into three parts: real datasets, benchmark, and synthetic datasets. Synthetic datasets could be generated to address specific scenario and conditions [29]. It is also used in developing and testing various algorithms in a real-time environment.

In large traffic networks, a benchmark datasets are generated based on algorithm simulation. The simulation of different attack situations in the traffic network could be performed while the benchmark dataset is developed. The real-life datasets are usually formed by collecting traffics within specific time period. It consists of normal (non-incident) and abnormal (incident) features. Figure 2 discusses the various datasets for traffic classification in the IoT and other networks. These datasets are used to evaluate the performance

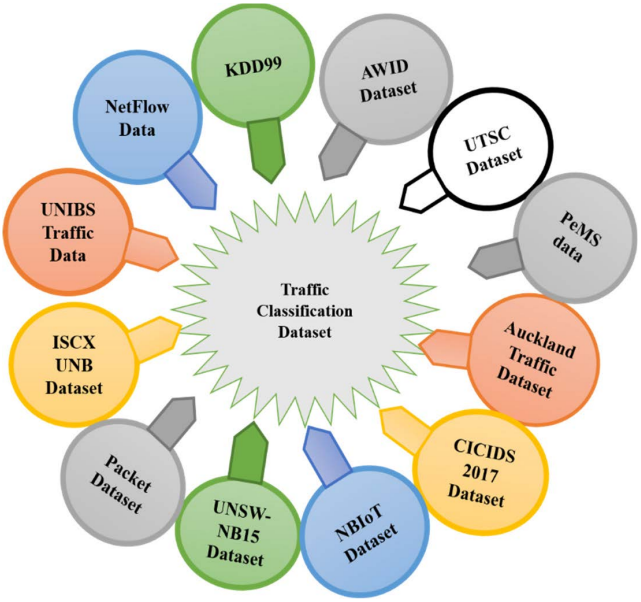

FIGURE 2. Traffic classification dataset.

of their algorithms. The technical details of the dataset are discussed below and are shown in Table 1.

1) NETFLOW DATA

The NetFlow dataset could be collected through the network switch or router as tracing the entry and exit of traffic flow can be easier at the network switch. Shafiq *et al.* [2] reported that Cisco NetFlow is considered as a unilateral packet sequence that has various features such as input port, output port, IP protocols and IP type, etc. The NetFlow data have two versions: compressed and processed version of the packet network. The architecture of NetFlow consists of various components such as collector, console, and exporter.

2) UNIBS TRAFFIC DATA

UNIBS is one of the most common datasets of traffic classification. It is developed by Prof. Gringoli and his team [29]. They collected traces using edge router at campus of the University of Berscia for three days. Then, they collected the data traffic using Tcpdump using malfunctioning router which is linked with the uplink of 100 Mps [30].

3) ISCX UNB DATASET

The ISCX dataset is developed using the concept of intrusion description and abstract details for various applications, protocols, and entities in the low-class network [26]. McHugh [31] collected data by using two different profiles such as α and β profiles. These profiles were used to form a new dataset in packet and bidirectional formats. α profile represents the abnormal or malicious behavior and β represents normal behavior performed by the network node. The dataset comprises of various network attacks, such as Botnet, DDoS, eavesdropping, Internet attack, etc.

4) PACKET DATASET

The applications are commonly used by researchers to generate network packet. Traces can capture packets that are transmitted and received using Libpcap and WinPCap at the physical interface [20]. In [32], Jacobson et al revealed that the most reliable applications to generate packet are commonly used in windows and tcpdum. An Ethernet frame named as Ethernet header (i.e., MAC) at physical layer of hundreds of payload bytes. In [20], Buczak and Guven revealed that the internet protocol of the payload could trace the packet using pcap interface.

5) UNSW-NB15 DATASET

The UNSW-NB15 dataset is developed by the Cyber Range Lab in Australia using IXIA PerfectStorm tool [26]. It consists of 100-GB raw data collected from the traffic network using tcpdump tool. More than 2.5 million raw data are segmented into different pcap files to analyze data record. The dataset with normal and abnormal (attacks or malicious) instances consists of training and testing parts, more than 175,000 and 82,000 records are found in the training and testing dataset, respectively [33], [34]. The UNSW-NB15 consists of various types of network attacks such as DoS, Worms, Generic, etc. along with features groups [34].

6) KDD99 DATASET

The KDDCup99 dataset was developed by the DARPA985-IDS in 1999 [35], [36]. The KDDCup 99 training dataset consists of over 4.9 million instances, in which normal and abnormal (attacks) are highlighted in 41 features. Also, it consists of 24 kinds of different types of attacks such as DoS, user to root, remote to local [16]. The testing dataset consists of more than 0.3 million samples. This dataset have been significantly applied to detect malicious behavior of traffic classification in the IoT network [37]. Reference [20], investigated whether the KDD could be used to extract useful information to obtain previous information. However, the imbalance training and testing dataset of KD99 leads to an inadequate performance for analyzing traffic classification.

In [38], Awid presented the NSLKDD dataset to overcome the imbalance issue of the KDD dataset. The author vanished the duplicate records of each instances, and resampled selected instances to highlight non-linear distributed issues. Reference [39] discussed that the KDD shows the entire process for obtaining information by input traffic data. They indicated that DM identified the particular part in the KDD process and data obtained from models. Tavallaee *et al.* [40] introduced the NSLKDD dataset to overcome various issues highlighted by [33].

7) NBIOT DATASET

This dataset provides a botnet dataset for traffic classification in the IoT network. It comprises of over 7.06 million instances obtained from the real traffic dataset. It contains malicious instances which are divided into ten attacks and are executed by two botnets: Bashlite and Mirai botnet [26], [41]. The

bashlite dataset consists of flooding, TCP/UDP, junk, etc. The mirai attacks consist of scan, syn, udp plain, and udp flooding [26]. Wireshark was used to record the traffic data using the traffic routers connected over the Wi-Fi network [42].

8) UTSC DATASET

The UTSC dataset is developed using two parts. One consists of various malware traffic from real-world traffic network instances from 2011 to 2015 by CTU researchers [43]. In UTSC dataset, the malware traffic dataset consists of various types such as Htbot, Miuref, Shifu, etc. The second dataset consists of ten different types of normal traffic obtained from simulating traffic network using IXIABPS [26]. The total size of the UTSC dataset is over 3.7 GB in pcap file. It consists of total 0.75 million records, at which malware data consist of over 0.4 million.

9) AUCKLAND DATASET

The Auckland II traffic dataset is commonly used for identifying traffic classification due to its accuracy. Auckland traffic data is obtained from GPS traces using DAG2 at the University of Auckland [44]. It consists of 85 traffic trace files which are collected from November 1999 to July 2000 [2].

To trace Auckland II dataset, a group of researchers from University of Auckland used the DAG3.2E card of 100 Mbps. They aim to identify and trace traffic at the router which is placed at border of University firewall. Nevertheless, the port numbers could identify the application type traces [2]. Peng *et al.* [45] used Auckland dataset to demonstrate the performance of their traffic classification model. Firstly, they gathered 8 types of applications obtain from traffic traces of Auckland dataset. Second, Peng *et al.* performed filtering on the traffic flow using different non-zeros packets to obtain traffic classification.

10) AWID DATASET

An Aegan Wi-Fi Intrusion dataset (AWID) comprises of traces data from the dedicated network 802.11 using a SOHO network in the Physic Research Lab [46]. This dataset consists of normal and abnormal instances, and some instances records were used for training and testing of dataset. The size of AWID dataset is around 935 MB that contain total of 1.795 million instances, in which over 1.63 million instances are normal traffic and over 0.16 million instances are abnormal traffic [26]. Reference [46] collected a dataset by running for one hour with attacks that last for only 15 minutes. Moreover, the number of attack instances in the training and testing dataset is about 162385 and 44858, respectively. The types of attack in AWID consists of impersonate, injection, and flooding.

11) CICIDS 2017

The CICIDS2017 dataset consists of the results of traffic network analysis using traffic label flows which are based on source and destination ports, Internet protocols, and time stamp. It also consists of various updated attacks, which are similar to real-world data (PCAPs) [47]. The Canadian Institute for Cybersecurity captured the data for about 5 days, from July 3, 2017, to July 7, 2017. They implemented various kinds of attacks such as DDoS, Web Attacks, Botnet, etc. in the data [47]. The CICIDS2017 dataset contains significant numbers of features and traffic, which could be used to detect anomalies [48].

B. EXTRACTION AND SELECTION OF FEATURES IN TRAFFIC CLASSIFICATION

The extraction and selection of features (ESF) plays a significant role in network traffic classification and identification. Without them, it is very difficult to identify and classify various classes in the traffic network. The selected features are directly related to the effectiveness of the traffic classification algorithms. Also, the number of extracted features could also affect the performance of traffic classification in terms of speed of classification and identification. Therefore, it is necessary to understand the concept of the ESF which could reduce the dimension of data and to develop the relationship between different features. The ESF method can be further classified into different types, such as filtering, wrapping, and embedding [49].

Recently, a few studies have adopted various learning-based techniques to improve the performance of traffic identification [2]. These methods could accurately identify traffic using different datasets in various traffic networks. However, researchers and practitioners could face imbalance traffic issues for classifying traffic in network identification. An imbalance class of traffic classification remains a critical issue in traffic identification. To overcome this problem, researchers from Electronics and Communication Technology background proposed various solutions in which the ESF plays a major role for identifying the traffic identification. Wasikowski and Chen [50] developed the various methods for traffic classification and then analyzed and compared with different metrics in terms of imbalance class issue. They also introduced various features such as signal noise and feature assessment in order to manage imbalance class of traffic classification. Similarly, Lim *et al.* [51] examined the features and selection for class distributions of traffic identification. Also, Peng *et al.* [45] investigated different features, which were used to evaluate traffic classification at the initial stage. They examined that the early features of traffic classification could obtain a large amount of packets at beginning stage of obtaining traffic identification. Moore *et al.* [52] introduced attribute selection algorithm for traffic classification. Moore *et al.* extracted different types of 248 statistical features based on traffic flow, and evaluated them in terms of network packet size and statistical features. These features could lead to obtain a better traffic classification results. Bernaille *et al.* [53] discussed the issues of the feature selections and considered packet size as a feature. They extracted various attributes from the packets and applied various models such as HMM and GMM to identify the traffic network [54].

Note that, one of the most important tasks in traffic classification is feature extraction by using the trace analysis. Recently, learning-based methods have been used to obtain the feature selection in traffic classifications. In this regard, Ding *et al.* [55] discussed the procedures for extracting features such as linear and non-linear features in detail. Also, Bennasar *et al.* [56] presented various non-linear techniques for extracting a better features for traffic classification. These techniques could provide a better features selection as compared to other methods. Zhang *et al.* [57] investigated the issues of the feature selection for traffic classifications. They developed a feature algorithm using a weighted symmetrical uncertainty (WSU) method and then select the stable features to identify the best feature using the WSU technique. Similarly, Chen *et al.* [58] revealed that the feature extraction could be used to obtain accurate network traffic classifications based on the time and location of features.

C. DECISION PROCESS (DP)

The DP plays an important role for obtaining the traffic classification. It relies on the extracting and selecting feature of traffic classification by employing ML algorithms or pattern matching technique. The ML algorithms are widely used for obtaining traffic classification and details of these algorithm can be found in Section 6. While the pattern matching (PM) is depend on the number of designated packets. The string matching algorithm could be used to compare with the string library in order to classify and identify traffic. However, the PM requires a larger computational time when processing the complex library and services.

D. VALIDATION PROCESS (VP)

The VP tests the outcomes of the previous traffic classification in order to determine the accuracy of traffic classification algorithms. To accomplish this task, first, we compare the values which are obtained from the original data with the experimental results. As a result, we can obtain the accuracy of the traffic classification method. Obtaining the collection of various categories within the original dataset remains a challenging issue. The ground truth collection method is widely used for labelling the traffic using the port collection and DPI tool. However, these methods could provide unreliable information and consume a large computational timing to process traffic labels. To overcome these issues, a new collection method based on heuristic analysis has been developed [59]. Employing these approaches could be beneficial in terms of the reliability of collecting data, but they are vulnerable to the larger traffic loads.

III. TRAFFIC CLASSIFICATION CRITERIA

This section discusses the various eligibility criteria's for obtaining a traffic classification and how to determine its effectiveness. Traffic classification criteria can be accomplished based on the classification effectiveness and classification performance as illustrated in Figure 3.

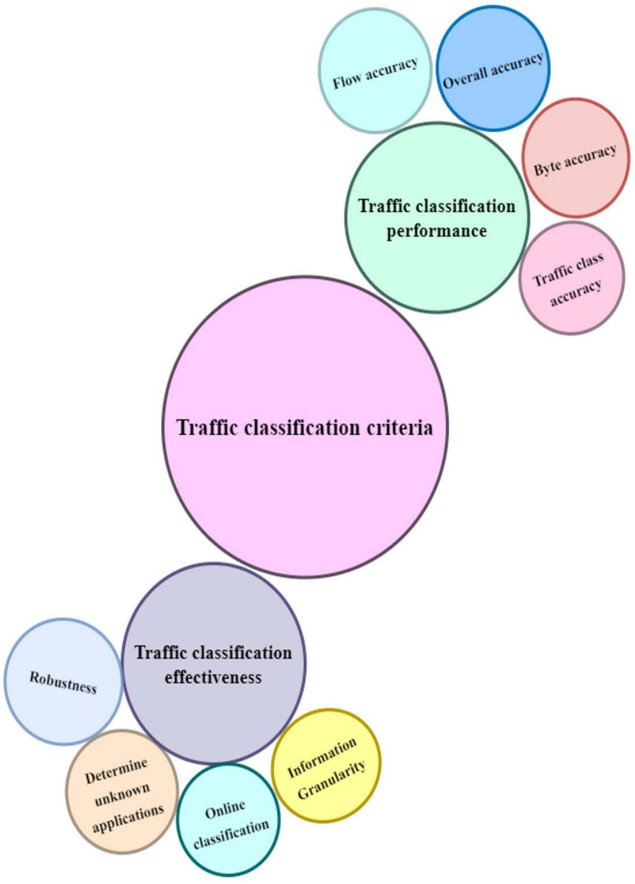

FIGURE 3. Traffic classification criteria.

A. TRAFFIC CLASSIFICATION EFFECTIVENESS

1) INFORMATION GRANULARITY

Information granularity (IG) plays a key role in determining the effectiveness of classification criteria's. The information obtained from granularity is depend on the type of granularity. The granularity could provide a better classification and the information obtained from the granularity is more reliable, accurate, and also provides enhanced data access. The traffic can be classified with different requirements as per the distinct criteria.

2) ONLINE CLASSIFICATION

An online classification (OC) could be used to assess the traffic classification algorithms in terms of real-time evaluation. The traffic network can update on the regular interval which enables the traffic classification methods to classify and identify the traffic online. Since the classification of the traffic network is online, therefore, it plays a key role for improving the network performance and detecting the malicious traffic nodes and activities. This can be accomplished by identifying the traffic class and category in a short period of time.

3) DETERMINE UNKNOWN APPLICATIONS

The traffic classification algorithms are commonly used to classify and identify label traffic within the training dataset, and to detect various new applications within data. Detecting

new applications can be further divided into various types known categories. When the traffic network environment is constantly changed and updated then the likelihood of appearing unknown traffic flow is higher. Therefore, it is essential to accurately identify and classify the unknown traffic which could lead to identify malicious traffic node and enhance the overall performance of the network.

4) ROBUSTNESS

The aim of traffic classification algorithms is to obtain a stable and reliable performance in a rapidly changing traffic network. It can provide a better classification accuracy by overcoming the various network issues such as packet delay, traffic loss, etc. Therefore, determine the robustness criteria plays an important role prior to designing and implementing the classification algorithms. Note that, the robustness of traffic classification can evaluate in terms of determining the universal features and whether the designed classification algorithm can provide a reliable performance in different traffic network.

B. TRAFFIC CLASSIFICATION PERFORMANCE

There are various criteria's which could be used to assess the effectiveness of the traffic classification methods. Researchers could use different performance metrics such as such as false positive, false negative, accuracy, etc., to measure the performance. This survey focuses on the identifying the accuracy criteria's for obtaining traffic classification methods.

1) TRAFFIC CLASS ACCURACY

The class accuracy (CA) directly related to traffic classification accuracy in terms of individual class. For instance, when the algorithm divides the network traffic into various categories such as HTTP, SMTP, etc., then the accuracy of these methods determine separately, which makes it more efficient to determine which traffic class is sensitive to the classification technique. The CA could be beneficial to identify merits and demerits of the classification algorithm.

2) BYTE ACCURACY

The byte accuracy indicates the number of bytes are correctly classify into the training dataset. It plays an important role when the dataset are imbalanced because the bytes are generated by the mice flows in the Internet. If the generated bytes by the smaller numbers of traffic flows which could be useful for a large portion of the dataset [60].

3) OVERALL ACCURACY

The overall accuracy is used to determine the number of instances which are accurately classified in the samples of training dataset.

4) FLOW ACCURACY

The flow accuracy employs in algorithms to identify and classify traffic flow such as correlation-based methods.

IV. TRAFFIC CLASSIFICATION TECHNIQUE

Traffic identification and network classification provide significant improvements to enhance the QoS and network security, and network traffic management. Exacting traffic identification can enhance network environment, network monitoring and network security. Network traffic operators and service providers can control the network performance, maintain and manage network resources. Also, service providers can find network growth and manage available network resources for applying on specific applications [2]. Figure 4 shows the various traffic classification techniques. The encrypted traffic classification is crucial for network security and is widely used to ensure data and network security. It also provides various technical support to improve QoS [61]. However, encryption techniques can make the detection of abnormal traffic even more difficult [62]. The significant increase in encrypted network traffic could limit the effectiveness of traffic classification techniques because the packet inspection techniques are unable to obtain the network information from the network traffic. For instance, most of the Internet traffic is associated with P2P applications, but the classification of the P2P traffic remains a complex task [26].

Hurley *et al.* [63] proposed the application growth of traffic day. The P2P applications consumes a large amount of bandwidth due to the bidirectional traffic flows. Moreover, different other application such as HTTP, FTP, etc. can consume large amount of bandwidth in the network. ISP also facing many challenges to provide efficient services to their customers. Such challenges are broadband quality, customer services, upstream bandwidth, etc. Mohammadi *et al.* [64] proposed a hybrid scheme to classify P2P traffic in IP network using genetic algorithms and neural networks. They showed that the P2P applications occupy 60% of the total available bandwidth. However, it's difficult for ISPs to achieve QoS and implement the network security and intrusion detection system for every traffic within the network. In particular, traditional traffic classification discussed the classification problems and identification of various applications to ensure network security from different perspectives. Generally, the IP based traffic classification consists of various inspection packets of TCP or UDP ports numbers, which indicates that these packets are either a port-based or a payload classification. Schulze and Mochalski [65] launched a survey on network traffic management worldwide. The P2P application can produce a more network traffic as compared with the other network applications such as online streaming, online gaming, messaging services, etc. The authors revealed that the web traffic gain significant attention due to access of social networking websites and file sharing policy. The process of the traffic classification techniques are highlighted in Figure 5.

A. PORT BASED TRAFFIC CLASSIFICATION

Port-based classification (PBC) is one of the commonly used technique for traffic classification associated with port

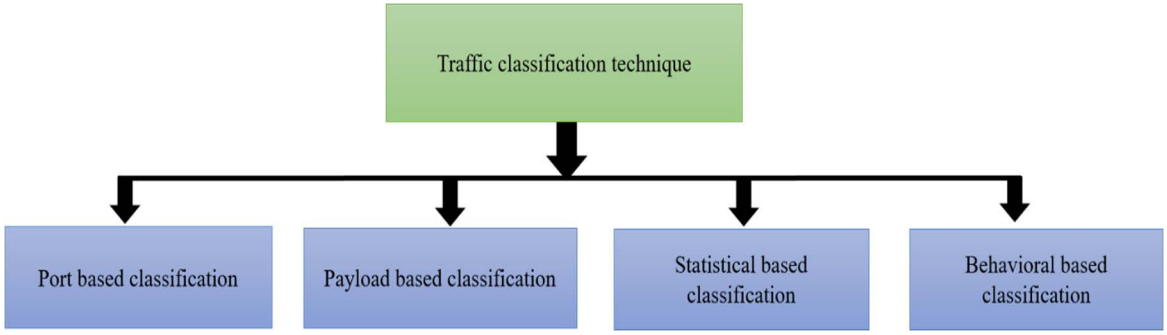

FIGURE 4. Block diagram of traffic classification.

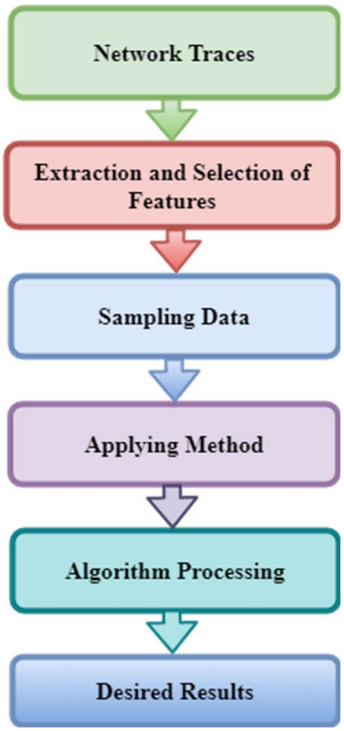

FIGURE 5. Traffic classification model.

number to applications [26], [66]. PBC examines the packet header and matching its inspection with TCP and UDP port number for register an application on the Internet. TCP and UDP generate the multiple flow connection using port numbers between public IP endpoints. The classification is to use well-reputed web traffic associated with TCP port 80. As this technique only checks the packet headers, it is fast when applying a light complexity calculation.

The port based technique plays a very important role to classify and identify network application in a huge traffic network. Nevertheless, it requires dynamic port number such as Napster, P2P, etc. [11]. The real video streamer port was developed for data transfer. However, Moore and Papagiannaki [67] revealed that the real video port does not get over 70% of port-based traffic identification and may not be able to provide a better classification accuracy.

Application developers become more intelligent to protect their applications from detecting system [26]. The latest advancement in the network technology provides access to usage of nonstandard applications. It allocates dynamic port number that deteriorate the performance of classification and makes its less productive for different applications. In such cases, the classifier causes numerous amount of false negative results. Also, some of the applications hide themselves behind well-reputed ports such as illegal applications use HTTP traffic over TCP port 80. As a result, the applications may be untraceable, which results in false positive rates of classifiers [26].

B. PAYLOAD BASED TRAFFIC CLASSIFICATION

Payload based traffic classification is also known as deep packet inspection (DPI). The packet and characteristics of network applications are analyzed signature features and applications of network traffic [2]. The DPI is especially designed for P2P applications of traffic classification used as dynamic numbers. Several methods have been proposed that studied an analyzing signature traffic, which can reduce 5% false-positive and false-negative for P2P traffic [11], [68]. Moore and Papagiannaki [67] proposed a hybrid method that utilized payload and port based methods to identify various network applications. Firstly, they used the classification number to determine the network flow. Then, they examined whether network flow contains signature or not. The proposed method classify 69% of internet traffic, and it obtained 79% of classification accuracy. This technique is not reliable for traffic classification due to its equipment cost for checking payload patterns. The intrusion detection systems are commonly used for payload classification for identifying malicious activity. The false-positive and false-negative results can be minimized by applying the payload classification to approximately 5% [68]. Liu *et al.* [69] proposed a method for mobile traffic classification using the extended labeled data (ELD). First, different traffic identification tasks were performed by SeverTag, payload distribution inspection and Random Forest. Then, ELD was used to identify mobile traffic using the encrypted payload. Yang *et al.* [70] proposed a payload based classification model that uses the concepts of handshake packets to classify encrypted traffic. They used the Bayesian neural network based classifier to

process handshake packets as inputs. Then, they used them to classify encrypted traffic. The authors claimed that the proposed model outperforms other traditional payload based classifiers.

Still, the payload based classification techniques have several demerits. It is difficult to examine encrypted packet contents at payload based classification technique. The packet evaluation with encrypted contents becomes very complex using payload based classification and most of traffic remains unable to classify [71]. Finsterbush *et al.* [14] launched a survey on payload based traffic classification techniques. The author discussed the performance of various DPI open-sources solutions such as OpenDPI, Hippie, and Libprotoident. The OpenDPI and Hippie are able to classify different protocols such as HTTP, SIP, and Oscar with 100% accuracy due to unencrypted traffic. Moore and Zuev [72] revealed that the payload based traffic classification is unable to accurately recognize traffic type and conditions due to variation in payload signatures.

In the payload classification, network privacy and regulations remain a critical issue. Nevertheless, the packets contents in the payload approach is examined thoroughly using a payload-based technique, which results in violation of regulations and privacy policies. Also, the payload technique is very complex and computationally expensive. The payload classification has several advantages over the port-based method, but it is still unable to perform better performance on high speed networks.

C. STATISTICAL BASED TRAFFIC CLASSIFICATION

The statistical-based classification is also known as the rational-based classification technique. This technique applies flow level measurement to identify the statistical features of traffic [73]. The packet arrival time, packet lengths, and traffic flow idle time are examples of statistical traffic [73]. They are useful for network classification and identification for differentiating application types in the network. In the statistical technique, the classifier uses ML and DM techniques to deal with various traffic patterns obtained from large datasets, causing a higher computational cost. The statistical classification relies on the flow information, and its effectiveness depends on the features extracted from the flow [26]. The statistical classification can overcome the limitations of payload techniques since they do not rely on network packets inspection contents. Therefore, it allows efficient classification of encrypted traffic [71]. The accuracy of statistical based classification can be improved by identifying the best features from feature extraction and selection techniques and trying to train and classify datasets using various ML methods. Several methods related to encrypted traffic classification have been proposed in recent years. Alshammari *et al.* [74] presented a method for the identification of VoIP encrypted traffic using the ML method. First, they applied different ML methods such as C5.0, AdaBoost, and Genetic programming to generate signatures for identifying encrypted traffic. The results show that the proposed method could significantly improve the performance of VoIP encrypted traffic. Muliukha *et al.* [75] proposed a method for classifying encrypted traffic using the ML technique. They classified the traffic generated from technical virtual connections and VPN traffic. In VPN traffic, they considered IP address, total number of packets, port to classify the traffic. They tested the classification of these technologies using the random forest algorithm. Obaidy *et al.* [76] proposed an encrypted traffic classification model based on ML methods for identifying social media applications such as Skype, WhatsApp, etc. First, they collected the data using Wireshark from end-user machines to generate the traffic for social media applications. Then, they used the feature selection based on the Wireshark tool to select 14 bidirectional traffic for obtaining better classification accuracy.

D. BEHAVIORAL BASED TRAFFIC CLASSIFICATION

Behavioral traffic classification is generally considered to analyze traffic pattern such as traffic volume, traffic shape, pick load, etc. [26]. Al khater and Overill [71] revealed the type of application running on the host. In the past, a few research works used experimental information, in which the numbers of distinct ports and protocols of transport layers were used to analyze the pattern of network traffic [11], [77]. Jin *et al.* [78] investigated that the behavioral based classification technique can analyze various information. They applied a graphical visualization scheme to examine the connection between endpoints. The results revealed that the client-server applications are discriminate to different P2P applications. Bermolen *et al.* [79] presented a method for specific classification of applications in networks. It obtained a low computational cost by applying the behavioral classification technique. Kohout *et al.* [80] used learning communication patterns to determine malware in HTTP encrypted data. They used the snapshots of individual user activities that use contextual information to compensate for the inconvenience caused by encryption. Then, they proposed statistical descriptors in terms of communication snapshots that can be used by various ML algorithms to analyze traffic data. Experimental results show that the proposed method can be used on a Hadoop cluster.

V. MACHINE-LEARNING METHODS FOR TRAFFIC CLASSIFICATION

Different ML methods could be used for intrusion traffic classification and identification. Methods for identifying traffic classifications using ML-based methods are illustrated in Figure 6. Table 2 shows the summary of traffic classification methods and their features and applications.

A. BAYESIAN NETWORK

Bayesian network is used to highlight the variables along with their relationships as Probabilistic Graphical Model (GM) [127]. The network design consists of continuous or discrete variable nodes, and the edges of the network demonstrates the connection between these nodes [2]. Buczak and

TABLE 1. Datasets of traffic classification.

Dataset	Nature of Attacks	Number of Instances	Network Environment	Year	Network Type
KDD99 [35], [36]	• DoS • R2L • U2R • Surveillance	Training dataset: 4.9 million. Testing dataset: 0.3 million.	Testbed network environment.	1999	Benchmark (IDS).
AWID [46]	• Impersonate • Injection • Flooding	Training dataset: 16.238 K Testing dataset: 44.8 K	SOHO Network.	2014	Benchmark.
Auckland Traffic Dataset [44]	• DoS • Other Network attacks	GPS traces: 85	DAG3.2E network card.	2020	Benchmark.
UTSC Dataset [43]	• Htbot • Miuref • Shifu	Total instances 0.75 million.	Using a real traffic network into two different parts.	2017	Real-world Network.
NBIoT Dataset [26], [41]	• TCP/UDP • Junk • Scan • Syn • UDP flooding • UDP plain	Total instances 7.06 million.	IoT network.	2018	Real-world Network.
UNSW-NB15 Dataset [34]	• DoS • Worms • Generic	Total instances 2.5 million.	Testbed network environment.	2015	Benchmark.
Packet Dataset [20]	• DoS • Web attack • Other Network attacks	Hundreds of payload bytes	Ethernet header	2016	Pcap interface.
ISCX UNB Dataset [31]	• DDoS • Botnet • Internet attack • Eavesdropping attack	Total instance: 2.76 million	Network Emulation.	2017	Benchmark.
UNIBS Traffic Data [29]	• Smtp • http • Others	Traffic Traces	Network router.	2015	Tcpdump.
NetFlow Data [2]	• Web attack • Eavesdropping attack.	Traffic traces can be collected through router.	Network router or switch.	2020	Compressed and preprocessed version
CICIDS2017 [47]	• DDoS • Web attack • Botnet	Total instance: 2,273,097 Total malicious instances: 557,646	Network router or switch.	2017	Real-world and PCAPs.

Guven [20] showed the small network nodes relies on the big node, in which each node stay at their random variable position and then conditional probability form. Maeda *et al.* [81] examined a TCP-level data of eighteen different locations at Dartmouth University Computer College for about four months. Maeda *et al.* extracted IRC network using filter layer in the whole data network. The Naive Bayes provided better result about 2.4% for low FNR about 15% for low FPR at the Dartmouth flow traffic in first part.

Hsieh *et al.* [82] gathered traffic data using network monitor. The dataset are very reliable and efficient for traffic classification. Though, the training and testing datasets consists of less amount of instances. Firstly, they demonstrated that the Bayesian Network could identify and classify traffic

TABLE 2. Summary of the traffic classification approaches.

Reference	Paper	Features	Classification Approaches	Year	Application
[20]	Buczak and Guven	The small network nodes relies on the big node, at which each node stay at their own random variable.		2016	
[81]	Maeda et al.	TCP-level data of eighteen different locations at Dartmouth University.		2019	TCP
[82]	Hsieh et al.	They revealed that the packet feature contents contains sufficient information for network traffic classification.	Bayesian Network	2019	TCP
[83]	Livadas et al.	To classify C2 flow of IRC botnet traffic.		2006	
[84]	Gao et al.	Employed the DARPA 1999 dataset using TCP/IP packets and define a set of attributes.		2019	TCP/IP
[85]	Shafiq et al.	They discussed the number of packets that are used at early stage for identifying and classifying traffic classifications.		2017	IP
[86]	Moore and Zuev	To categorize and classify Internet traffic classification in terms using various applications.	Naive Bayes	2005	IP
[87]	Park et al.	Used feature selection technique that relies on Genetic algorithm for obtaining traffic classification.		2006	
[88]	William et al.	They evaluated the computational speed of various algorithms such as C4.5 Decision Tree, Bayesian Network, etc.	Decision Tree	2006	
[89]	Zhou et al.	Separated the payload and port-based identification techniques to assess various network activity such as QoS, security and privacy.		2011	
[90]	Cui et al.	Estimated the cyberattack using Naive Bayes algorithm based on statistical features of data and CDF.	Artificial Neural Network	2019	TCP/IP
[91]	Bivens et al.	They demonstrated that the neural network could detect efficient malicious node activity within the network.		2002	
[8]	Erman et al.	They clustered the flows from 64K unlabeled flows for network traffic classification.		2007	
[92]	Mishra et al.	Classify and maps attacks corresponding to each attack. Also, discussed tools and future directions.	K-means Clustering	2018	
[93]	Erman et al.	They identified and differentiate among web and P2P traffic network. Also, determine the performance in terms of unidirectional trace packet.		2007	P2P, FTP
[94]	Kant and Mahajan	Presented an outlier detection technique by combining the K-mean and PSO algorithms.		2019	Healthcare

TABLE 2. *(Continued.)* **Summary of the traffic classification approaches.**

[95]	Jian and Pamula	They altered the K-means classifiers and used them for obtaining data pattern. And, designed a max heap which depends on number of cluster.		2019	Intrusion detection system.
[96]	Gauci et al.	Discussed the utilization of reinforcement learning method by Facebook such as push notifications	Deep Reinforcement Learning	2020	
[97]	Kalashnikov et al.	Train a neural network Q-function by employing over 1.2M parameters to perform multiple tasks.		2020	
[98]	Wagner et al.	They obtained a real-world data using Internet services and Flame tools.	Support Vector Machine	2011	POP spams, Scan.
[99]	Yang et al.	They designed a classifier which is used for P2P traffic classification based on network traffic.		2008	P2P, PPLive.
[100]	Gyanchandani et al.	Discussed various rules for using fuzzy logic.	Fuzzy Logic	2012	
[92]	Mishra et al.	They revealed that the fuzzy logic system requires simulation test before implementing practically.		2018	IDS.
[101]	Mirzakhanov	Discussed a case study of fuzzy logic framework in terms of ML and DM techniques.		2020	
[102]	Raja and Ramaiah	Applied the knowledge obtained from fuzzy sets to detect intrusions.		2016	IDS.
[103]	Xie et al.	They discussed the non-cooperative game theory, in which the players are compete with each other to form their own strategies	Game Theory based Reinforcement Learning	2018	
[104]	McLaughlin et al.	They learned the malware features from the network based on the raw opcode sequence (ROPS).		2017	IDS.
[105]	Wang et al.	They designed the CNN-based traffic classifier for representation learning.		2017	P2P.
[106]	Wang et al.	Developed an end-to-end encrypted traffic classification model using one-dimensional CNN	Deep learning	2017	Cyberspace security, IDS.
[107]	Rezaei and Liu	Used time series features as input of the sampled packets		2018	
[108]	Wang et al.	Developed a real-time traffic classification method based on the parallelized CNN model		2019	
[109]	Zhou et al.	Used the spatial pyramid pooling (SPP) framework for classifying traffic.		2018	
[110]	Salman et al.	Develop a DL method to classify traffic based on various QoS and network policies.		2020	Network security.

TABLE 3. Traffic obfuscation techniques.

Traffic obfuscation method	References	Approach	Year	Features
BuFLO	[111]	Mutation	2012	Traffic sharper could fixed packets size.
Obfs2	[112]	Encryption	2012	Full message encryption (TOR plugin)
Flash Proxies	[113]	Tunneling	2013	It uses web sockets based on proxies.
Freewave	[114]	Morphing	2013	VoIP data transfer.
Meek	[115]	Tunneling	2014	Employ various domain names.
Blindspot	[116]	Steganography	2014	Hide the data into the social network.
FTE	[117]	Morphing	2013	Used for format transform encryption
Deepflow	[118]	Steganography	2014	Hide Internet traffic into P2P.
Facet	[119]	Steganography	2014	Hide a video streaming traffic in Skype communication tool
eMule	[120]	Encryption	2007	Determine the encrypted socket
Liberate	[121]	Mutation	2017	Insertion of packet, packet recorder, classification flush
Muffler	[122]	Morphing	2015	Extract distribution, at which the stream can morph.
Covertcast	[123]	Steganography	2016	It hides the web page materials into images.
Tamaraw	[124]	Morphing	2014	Different time interval of uplink and downlink of packets.
Deltashaper	[125]	Steganography	2017	Hide images data of the Skype video call.
Skypeline	[126]	Morphing	2016	Used steganography to hide information in VoIP communication.

network which are relies on network statistics features. Then, they revealed that the packet feature contents contains sufficient information for network traffic classification and identifications. Livadas *et al.* [83] examined various ML-based technique in order to classify C2 flow of IRC botnet traffic. Livadas *et al.* divided the whole process into two stages. Firstly, they study the characteristics of IRC and non-IRC botnet traffic flow. Secondly, they differentiate IRC traffic flow and botnet in a traffic network. Auld *et al.* [128] proposed a ML-based technique for traffic classification. It achieves a robust performance and accuracy without using application information. Auld *et al.* used traffic feature that are derived from packet content in order to obtain a better classifications. They demonstrated that the accuracy for traffic classification is better than Naive Bayes technique, and it could classify traffic flow about 99% and 95% of accuracy in testing and testing dataset, respectively. Gao *et al.* [84] introduced an intrusion detection system using ML method. Gao *et al.* employed the DARPA 1999 dataset using TCP/IP packets and define a set of attributes in their proposed method. They used a different threshold value to obtain a better classification accuracy.

B. NAIVE BAYES

Naive Bayes are the robust classification technique used as a ML-based classifier [129]. As it relies on Bayes Network theorem, it is not very difficult to develop large dataset for traffic classifications. And, the Naive Bayes is reliable and efficient technique for solving complex traffic identification and classifications. Naive Bayes approach has various classifiers which could be used to use attributes in precise network class. Reference [130] studied various traffic classification technique such as ML classifiers. Several ML-based technique have been used to obtain a better traffic classifications. The results obtained from simulation shows that the Naive Bayes could obtain a better traffic classification and identifications in any traffic network [2].

FIGURE 6. ML techniques for traffic classification.

Shafiq *et al.* [85] discussed the number of packets that are used at early stage for identifying and classifying traffic classifications. Firstly, they used the different type of packet size for extract the information. Then, they perform the mutual analysis to recognize the common connection among the data packets. Secondly, they used various ML-based technique with crossover identification method. Then, they performed various test to test and identify the packets that could be used for obtaining traffic identification and classification at early stage. They used two different dataset at early stage of network classification and compare them to obtain a better performance of the proposed system. Moore and Zuev [86] presented supervised ML-based Naive Bayes technique to categorize and classify Internet traffic classification in terms using various applications. They used the traffic flow dataset which are manually classified to obtain a better evaluation. There are 248 full features were applied to train the classifier. Secondly, the selected traffic are used for Internet applications that are formed into various groups for obtaining traffic classifications.

C. DECISION TREE

The decision tree technique is a category of supervised ML-based algorithms. It could accept input and output variables in the form of continuous and categorical types. The decision tree consists of many leaves that has several branches, at which you can represent traffic classifications. There are two ways to build the decision tree such as C4.5 [131] and ID3 [132] ML-based classifiers. These tree algorithms were designed on the decision tree using the training dataset by applying the entropy concept.

The decision tree consists of two types such continuous variable and categorical variable [2]. These terminologies are connected to decision trees in terms of root node, splitting, terminal node, and decision node, etc. The decision tree technique is used to determine the relationship among variables and form new variables, which could identify and classify the target variable efficiently. The decision tree requires less amount of traffic dataset, and it can apply in numerical and categorical variables [2].

Park *et al.* [87] presented a feature selection approach that relies on a Genetic algorithm for obtaining traffic classification. The authors tested and demonstrated the efficiency of the proposed method using three different classifiers such as Decision Tree J48, the Naive Bayes classifier with Kernel Estimation (NBKE), and the Reduce Error Pruning Tree classifier. The numerical results obtained from simulation reveals that the classification results of decision trees are more reliable and accurate than the NBKE technique. William *et al.* [88] studied the performance of the various ML-based traffic classification method such as C4.5 Decision Tree, Bayesian Network, etc. They calculate the computational speed of each algorithm by the processing of classification number per second, and the amount of time required to develop the classification model. William *et al.* tested the model using public NLANR traces and selected the classification features using a correlation-based feature selection model. The result shows that the performance of the Decision Tree algorithm was the best among other techniques. It obtains maximum accuracy of about 5.4700K classification per second by using any feature set.

Shafiq *et al.* [133] introduced a feature selection based method named as weighted mutual information (WMI) technique. Secondly, they introduced a feature selection methodology to select the best features which provide better accuracy. They demonstrated the effectiveness of the proposed method using five different ML-based classifiers. The simulation result shows that the presented method can select the select best features. Also, the Decision Tree C4.5 obtains a better accuracy among other ML-based techniques and produces the best classification results.

D. ARTIFICAL NEURAL NETWORK

An artificial neural network (ANN) technique is one of the most prominent ML-based techniques and considered a very reliable technique for conventional regression and statistical data modeling [134], in which the relationship between input and output data are model by ANN tools [2]. The main advantage of ANN technique is the robust processing on a large scale parallel implementation that significantly contributes and fulfill the need of research in ML-based technique [135]. The ANN gains significant attention from the researcher from Computer and Network Technology background due to its ability to process artificial neuron that are able to perform various computational process on the applied inputs. When a support vector machine (SVM) method was developed at that time the ANN model gains so much popularity among

other ML-based techniques. However, both ANN and SVM techniques take a large amount of time to process applied inputs.

Sun *et al.* [136] designed a model for determining exact collection of sample of traffic named as DHTCP. They demonstrated that their developed model is able to collect traffic sample on host user based on applied information. Sun *et al.* used probabilistic neural network based on DHTCP dataset for traffic classification. Finally, they employed different statistical features model to identify traffic. Zhou *et al.* [89] studied feed-forward neural network in order to obtain an efficient Internet traffic classifications model. They separated the payload and port-based identification techniques to assess various network activity such as QoS, security and privacy. They revealed that the fast correlation for obtaining feature selection could obtain a better performance than the neural network technique [2]. They scale dataset from 0%-100% in order to obtain an accurate classification results.

Cui *et al.* [90] developed a ML-based anomaly detection (MLAD) technique. Firstly, they used the load forecast obtained from neural network that are used to reconstruct benchmark and scaling data using K-means clustering method. Secondly, Cui *et al.* estimated the cyberattack using Naive Bayes algorithm based on statistical features of data and cumulative distribution function (CDF). The results shows that the proposed method could detect cyberattacks with higher accuracy but some network attacks were not effectively detected and obtained an accuracy of about 76%. Bivens *et al.* [91] investigate intrusion detection system using neural network. They demonstrated that the neural network could detect efficient malicious node activity within the network.

E. K-MEANS CLUSTERING

K-means clustering technique is one of the popular unsupervised ML-based algorithm. It can identify unlabeled data in different clusters. In order to implement K-means clustering, it requires two important parameters to process: the dataset and the number of clusters. When the cluster quantity is K, then the K-means clustering algorithm is used to overcome clustering issue into three folds: to initialize the K cluster, use of distance function at each network node closet to center node, and assign a new centroid by considering a current node and halt the classifier [103].

Some researchers uses K-means clustering as a classifier in order to obtain normal and malicious behavior of node within the traffic network while other uses K-means clustering algorithm to separate outliers and generates robust dataset for ML-based algorithm [2]. Erman *et al.* [8] employed a K-means clustering algorithm for network traffic classification. Firstly, they clustered the flows from 64K unlabeled flows. Then, a fixed amount of flow are labeled in each formed cluster. The results show that the presented method obtain about 94% of accuracy in labeled flows. Mishra *et al.* [92] investigated and analyzed ML-based methods in detail. These methods were used to determine the issues of ML-based techniques when detecting intrusion activity of the user node. Mishra *et al.* classify and maps attacks corresponding to each attack. Also, they discussed the various tools and future directions for detection of attacks using ML methods.

Erman *et al.* [93] introduced a method for identifying and differentiate among web and P2P traffic network. They utilize the clustering ML-based technique with the comprehensive demonstration of K-means clustering algorithm. Also, they evaluated the performance in terms of unidirectional trace packet. Kant and Mahajan [94] proposed outlier detection technique using the combination of K-mean algorithm and PSO algorithm. They demonstrate the performance of the proposed method using time-series dataset. The result shows that the presented method could obtain a better outlier detection. Also, they demonstrated that the can be used in different applications such as traffic classification and detection, medical, etc. Jian and Pamula [95] introduced two-step anomaly detection technique based on clustering algorithm. Firstly, they altered the K-means classifiers and then used them for obtaining data pattern. Secondly, they designed a max heap that depends on number of cluster. The numerical results shows that the proposed method obtained a better classification and identification. However, they used the simulated and iris datasets that are commonly used to identify malicious behavior of the node.

F. DEEP REINFORECEMENT LEARNING

The deep reinforcement learning (DRL) algorithm has capability to solve complex problems in different applications such as Physics, Computer Science, and Engineering. It could be used without requiring an input to solve the mathematical model. The DRL uses the combination of deep learning and reinforcement learning in order to provide robust results of various ML-based algorithms. The DRL overcomes the long-term learning problems, and obtains a robust results in different playing games due to its ability to perform variety of tasks [137]. However, the DRL has some demerits such as a low convergence rate and inability to solve complex dataset [2]. Gauci *et al.* [96] discussed the utilization of reinforcement learning method by Facebook such as push notifications and using fastest video uploading and downloading using smart prefetching. Kalashnikov *et al.* [97] studied a vision-based learning technique which trains a neural network Q-function by utilizing over 1.2M parameters to perform multiple tasks.

G. SUPPORT VECTOR MACHINE

The support vector machine (SVM) is one of the robust ML-based techniques. It can identify and classify Internet traffic, and classification of large amount of traffic data [138]. It is used for regression and classification, which relies on the separation of hyperplane. For instances, the SVM technique is reliable and efficient when the number of sample instances are lower and the number of features are higher [2]. Buczak and Guven [20] discussed the two different classes that are

not separable with slack variables which are added and then determined the cost for the data samples. Buczak and Guven also discussed the quadratic optimization in terms of practical running time.

Yang *et al.* [99] introduced a P2P network classification technique using SVM. They designed a classifier that is used for P2P traffic classification based on network traffic. Also, they considered other network traffic such as P2P, PPLive, Skype, MSN, etc. Yang *et al.* demonstrated that labeling traffic samples and define network attributes in order to select and classify traffic [2]. Wanger *et al.* [98] proposed an evaluation technique for NetFlow records. They applied the temporal technique to the ML method. Wanger *et al.* obtained a real-world data using Internet services and Flame tools, and other services such as pop spams, scans, etc. The simulation shows that the proposed SVM method obtained a better performance on various kinds of attacks that are reported with an accuracy of about 94%.

Traffic classification plays a significant role for detecting malicious nodes, managing security and privacy issues of network, and network traffic management. It is necessary to discuss every aspect of network traffic classifications and provide solutions to overcome the classification issues using ML-based technique.

H. FUZZY LOGIC

Fuzzy logic is a problem-solving methodology and able to deal with different kinds of numerical data and linguistic knowledge. It controls the complex system without requiring prior knowledge of its mathematical model. Fuzzy logic differentiates from other traditional logic techniques that do not require true or false, on or off, etc. In fuzzy logic, a statement can assume any real number between 0 and 1that represents the degree of truth. Fuzzy logic was introduced by Prof Lotfi A. Zadeh of the University of California at Berkeley.

Fuzzy logic could infer the features and properties of the neural network. Neuro-fuzzy is one of the robust techniques for detecting the malicious activity of network nodes. Gyanchandani *et al.* [100] discussed various rules which are four folds: (i) Fuzzy logic can combine the input using other sources; (ii) Measures used by IDS have some fuzzy features, (iii) More alert features of IDS are fuzzy, and (iv) Fuzzy logic rules could be modified based on security applications. Mishra *et al.* [92] investigated that the fuzzy logic approach is unable to detect major attacks. Although, the fuzzy logic could obtain better performance when applying with other ML-based classifiers. The fuzzy logic mainly is used to correlate with the intrusion detection system. Raja and Ramaiah [102] proposed the fuzzy logic-based model for detecting intrusion into the cloud. They applied the knowledge obtained from fuzzy sets to detect intrusions. They tested the performance of the proposed model on different datasets.

Mirzakhanov [101] proposed a case study of fuzzy logic framework in terms of ML and DM techniques. They studied fuzzy logic and compared the quality and quantity difference between them, then, the fuzzy logic methods could perform better than non-fuzzy methods. Then, they presented an association rule mining (ARM) technique, which is cluster based technique that provides fusion of clustering. The experimental results obtained through various real-time datasets provides effective results. Also, [92] discussed some demerits of fuzzy logic, in which the fuzzy logic system requires robust tuning and simulation test before implementing practically. Then, they highlighted the challenges when designing and developing a model using fuzzy logic as compared to other ML-based solutions.

I. GAME THEORY BASED REINFORCEMENT LEARNING

The game theory based reinforcement technique is considered as the mathematical model which relies on decision making and strategic rational. The game theory consists of various components such as payoffs, players, strategies [139]. Also, it requires the number of players and strategy to process operation. In the game theory technique, the decision and strategy makers use the payoff or utility functions to identify the best strategies.

There are various types of game theory algorithms such as cooperative and non-cooperative game theory techniques. In the cooperative game theory technique, most of the players cooperate together and form various associations. This technique is based on decision making and to cooperate and form strategies together. Although, in the non-cooperative game theory, the players could compete with each other to form their own strategy [103]. The players did not communicate and know the strategy with each other. However, the sets obtained from the players reveal that the ending of the play using specific strategies. Therefore, reinforcement learning is mainly used to decide and select optimum strategies.

J. DEEP LEARNING METHODS

The deep learning (DL) methods have been used successfully in various filed such as computer vision, image and language processing, speech and pattern recognition, etc. The DL model has been emerging from communication technology to traffic classification and identification in recent years [140], [106]. The convolution neural network (CNN) is a type of DL model, which is used to facilitate imaging applications. The residual neural network is the part of CNN architecture, which consists of skip connections to overcome the gradient issues.

Several deep learning models have been proposed to classify traffic in recent years. McLaughlin *et al.* [104] proposed a deep CNN method for android malware detection based on the raw opcode sequence (ROPS). They learned the malware features from the network based on the ROPS. McLaughlin *et al.* claimed that the proposed model obtained a better performance than the n-gram based classification model. Wang [141] proposed a DL model to classify and identify traffic by considering 1000 bytes in each flow. Wang *et al.* [105] proposed a CNN-based traffic classification model for representation learning by considering traffic data as images. They determined the best traffic features

among other layers using various experiments. Similarly, Wang *et al.* [106] proposed an end-to-end encrypted traffic classification model using one-dimensional CNN. They integrated an end-to-end model with features extraction and selection for learning the relationship between raw input and encrypted output. Rezaei and Liu [107] proposed a CNN model that takes the time series features as input of the sampled packets. They developed a new model based on the learned weights that consist of a small labeled dataset. Martin *et al.* [142] proposed a model to classify traffic using the statistical features based on the CNN and recurrent neural network (RNN) techniques. The results show that the proposed model obtained better performance than other methods without requiring any Engineering features. Wang *et al.* [108] proposed a real-time traffic classification method based on the parallelized CNN model. They applied the spark and spark streaming platforms to model the requirements of the real-time classification of network traffic. Zhou *et al.* [109] proposed a traffic classification model based on the CNN using the spatial pyramid pooling (SPP) framework. They used the LeNet-5 model to replace the max pooling with the SPP. Salman *et al.* [110] applied a DL model to classify traffic based on various QoS and network policies. The authors claimed that the proposed model outperforms the previous model in terms of allowing the traffic classification at different granularity.

VI. OBFUSCATION TRAFFIC CLASSIFICATION

The traffic classification obfuscation techniques can be employed by attackers to attack the network without being detected by the intrusion detection system (IDS). We comprehensively discuss the obfuscation techniques, which could help to design a better classifier. In this regards, several methods have been proposed such as Iwai *et al.* [143] proposed a ML based adaptive identification method and identified the unknown traffic flow using the trained classifier. They tested the performance of the presented method based on existing obfuscation techniques such as direct target sampling [144] and tamaraw [145].

We reviewed ML techniques for traffic classification in earlier section. We will review methods for traffic obfuscation, which may affect the traffic characteristics. The obfuscation traffic classification techniques can be further classified as encryption, steganography, tunneling, mutation, morphing, and layer obfuscation. Table 3 shows the summary of some of the traffic obfuscation techniques and its approaches.

A. ENCRYPTION

The Internet applications depend on the user private and confidential information which needs to prevent them from any malicious activity such as an attacker modify the original information, copy useful information without any permission and disclose information to illegitimate users. Therefore, the traffic encryption could be adopted over the Internet to hide the user's useful information. In the traffic classification scenarios, the encryption mechanism can hide the signature which could be used for traffic packet-based inspection classification techniques. Also, some encryption algorithms require fixed length which could the affect length and size. Several ML-based classification methods were demonstrated the effectiveness of their proposed method by classifying and identifying encrypted Internet traffic based on the interval time and network packet size [146], [147].

B. TUNNELING

The traffic encryption do not ensure the total privacy, therefore, the tunnel protocols could be used to overcome this issue. The tunneling protocols could be used to hide the metadata connection and to ensure user policy. The famous tunneling service could be used to determine the virtual private network (VPN). The VPN depends on various protocols such as IKE, SSL, etc. Generally, VPN acts as a tunnel between the client and server, and the server is able to transmit the packets to its destination. Also, the VPN user (client) could encrypt the data prior to sending it to destination. A few works have been proposed which aims to classify VPN traffic based on the traffic class and type [106], [148].

C. LAYER OBFUSCATION

The traffic classification in the wireless and sensing network could be used to determine leak side-channel information and network signal pattern. Therefore, it is necessary for obfuscation traffic classification to apply them in order to use for land networks such as morphing, padding [149], and other approaches [150]. In this regard, Zhang *et al.* [151] proposed a traffic model which aims to develop media access control (MAC) interfaces and it scheduled packets over these interfaces, therefore, redevelop the features of packets based on each interface.

D. STEGANOGRAPHY

The steganography is a process containing confidential data in visual domain. It aims to hide the confidential data into packets, and send them through the network. Recently, a few steganography related works have been proposed which aims to ensure the untraceable of various protocols [152], [153]. Some steganography methods such as Deepflow could be used to hide the TOR in the P2P traffic. The Deepflow hides the unknown traffic in the P2P network using the steganography [118]. In particular, the deepflow node could be used to connect the PPS stream and it works as a client of PPS and transmit the data in the form of video packets and these packets are reach to their destination via PPStream nodes. Facet is another commonly used method of steganography obfuscation method [119], which aims to hide the video traffic of video communication tool such as Skype. First, the facet could be used to generate a message to the server giving the URL address of the video, which it wants to see. The server downloads the selected video and the content is transmitted via a Skype video platform to the facet client.

E. MUTATION

The traffic classification tunneling could use to hide the information contain in the network packet payload. The traffic classification could be obtained using the statistical flow features by considering the time interval and size of packet, and features can be modified based on traffic mutation mechanism. In this regard, the padding technique could be used to hide the network packet detail.

Linear padding: It consists of passing the traffic packet based as the expression below.

$$i\,(m) = \frac{*m}{n} * n. \tag{1}$$

where n denotes the system parameter and $\frac{*m}{n}$ represent the ceiling, and the length of packets are multiplication of n.

Exponent padding: The padding packets can be represented into the exponential form using the below equation.

$$i\,(m) = \min 2^{log_2(m)}. \tag{2}$$

F. MORPHING

The morphing technique could be used to divert the classifier based on the classifying the traffic application. It's widely used to avoid censorship issues of application protocols and make them legitimate in various applications. Wright *et al.* [154] presented a traffic morphing approach to obfuscation traffic analysis. They used to morph the one class traffic into another class traffic by using convex optimization method. They also assessed the performance of the present method against other classifiers such as Web traffic [155] and VoIP [156].

Wang *et al.* [157] revealed that the TOR traffic could be easily detection after obfuscating using two different variants such as format transforming encryption and obfsproxy. The TOR traffic has been used to overcome the issues in [158], [159]. The stegotorus is considered as a TOR plugin which aims to detect and unblock the TOR traffic. It employs steganography to form a TOR traffic which considered as the traffic which developed by another software [159].

VII. KEY FINDINGS, LIMITATIONS, AND RECOMMENDATIONS

This section highlights the key findings, limitations and recommendations after reviewing existing methods for employing ML based traffic classification as illustrate in Figure 7.

A. KEY FINDINGS OF TRAFFIC CLASSIFICATION

This section discusses the key findings for employing ML based traffic classifications.

1) DATA COLLECTION

The data which are used for obtaining traffic classification in most of the reviewed papers were not updated. In the recent years, traffic classification continues to evolve with latest trends and technologies such as new traffic devices and applications. Consequently, the collection of Internet traffic classification based on latest trends is necessary to overcome the issue. Most of the datasets were labelled using deep packet inspection technique which were unable to integrate with port-based labeling with dynamic port allocation.

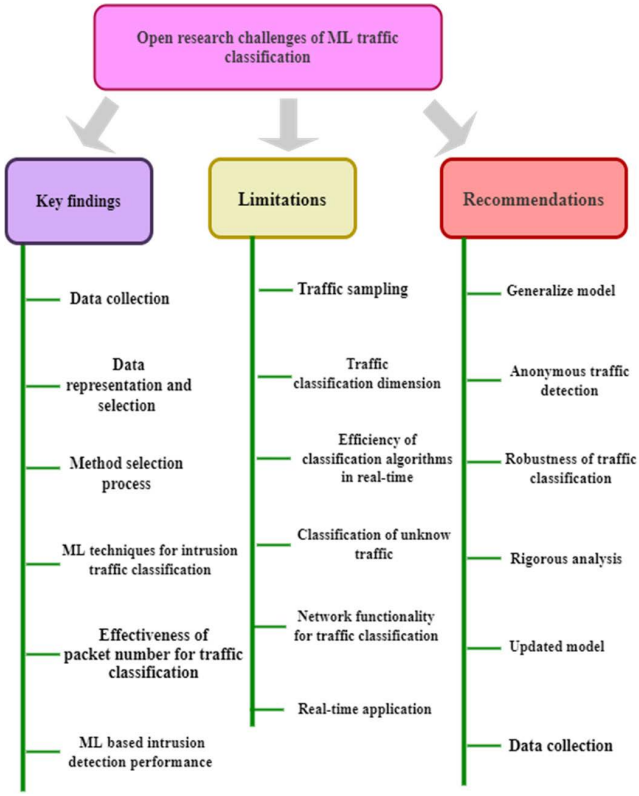

FIGURE 7. Open research challenges of traffic classification.

2) DATA REPRESENTATION AND SELECTION

Several methods have been proposed for traffic classification in recent past years. Moore *et al.* [52] revealed that various feature sets have relied on network packets, TCP flags, port numbers, and different IP addresses. While the other techniques use the domain name and used protocol request contents. Some researchers applied the datasets for detecting anomalies and identifying the intrusion traffic classification using ML algorithms. They investigated the misuse detection of cyber security and identifying traffic classification [160]. Most of the researchers used the available public datasets for validating the performance of their proposed algorithms, but designing an effective algorithm that employed these datasets is a difficult task. Therefore, it is necessary to develop a new dataset while reusing the available public datasets. Also, some authors use traffic trace while designing an ML-based traffic classification. Though, these traces do not contain accurate information related to traffic classification. Therefore, a new framework is necessary for collecting accurate datasets and effectively applied in various applications.

3) METHOD SELECTION PROCESS

The selection of method plays a significant role in obtaining the best performance of the algorithms. Several types of ML techniques are available and each of them has an advantage

and disadvantage. The aim of traffic classification is to obtain a better QoS and improve the privacy and security of the network. In the literature review, different network protocols, approaches, and application have been considered in order to select the best ML technique for traffic classification.

4) ML TECHNIQUES FOR INTRUSION TRAFFIC CLASSIFICATION

Machine learning (ML) methods play an important role in identifying traffic classification as discussed in section 5. They are useful for obtaining traffic classification based on various schemes such as statistical and entropy (decision), etc. Note that the training datasets have effective features and statistical properties. Therefore, it is necessary to determine whether the proposed model works with the online and offline datasets. We have found from the literature review that no study had proposed a robust ML-based classifier for traffic classification. Therefore, there is a need to develop a new ML based classifier and effective datasets should be used for obtaining a better traffic classification.

5) EFFECTIVENESS OF PACKET NUMBER FOR TRAFFIC CLASSIFICATION

Traffic classification plays a crucial role for network application and intrusion detection identification. It identifies and classifies the unknown traffic classes on the entire network. A few research works related to traffic classification using the ML method were proposed. Different approaches were presented to identify traffic classification, in which the early-stage traffic classification technique is popular among those methods. Several studies found that the early network packet up to ten is sufficient and effective for obtaining traffic classification. Some demonstrated that up to six early packets are enough for accurate traffic classification. Early-stage classification is still at the beginning stage, and it needs to determine how many network packets could be used for traffic classification through investigation and review. Researchers should also thoroughly examine and study feature extraction and selection methods. It needs a comprehensive review to enhance the performance of existing approaches by designing an effective feature set for traffic classification at an early stage.

6) ML BASED INTRUSION DETECTION PERFORMANCE

An effective dataset plays an important role for identifying a robust and accurate network-based anomaly detection. Several researchers used various datasets such as NetFlow data, packet dataset, and KDD dataset and applied various ML-based methods to identify effective traffic. Their studies did not propose appropriate methods for anomaly intrusion detection as they just assess the effectiveness of the proposed methods using anomaly datasets. Therefore, it is necessary to develop a new and robust anomaly-based intrusion detection system, which could evaluate the performance of these ML-based methods and identify the best algorithm among them. We have found from the literature review that several studies have employed the classical ML-based methods for detecting anomalies, but they did not propose a new ML method. For example, several authors studied the decision tree, deep reinforcement learning (DRL), fuzzy, logic, and game theory methods. They are effective ML-based methods for detecting anomalies, but a few studies have applied them. Therefore, a need for a new ML-based algorithm arises, which comprises of these ML-based algorithms, such as decision tree, deep reinforcement learning, and so on.

B. LIMITATIONS OF ML FOR TRAFFIC CLASSIFICATION

In this section, we will discuss the challenges and limitations that researchers may face when applying the ML technique for traffic classification.

1) TRAFFIC SAMPLING

The main challenges of traffic classification are that it hides the features of traffic classification applications due to the high speed requirements of the traffic network. It's not reliable to extract features from network packet which requires high speed network. Therefore, traffic sampling could be used to overcome this problem. It could transform different characteristics and features of the traffic, but requires a large computational time to process these characteristics and features, which may resulted in a lower accuracy of traffic classification techniques.

2) TRAFFIC CLASSIFICATION DIMENSION

The weight of traffic classification must be smaller when they are applied in real-time applications. To determine the traffic classification, data representation and selection must be analyzed carefully in terms of time and computational processing that are required to process the traffic classifier. Under this condition, several factors need to be considered such as algorithm complexity, memory space, and computational timing in order to design a better classifier. However, it may cause several issues, such as delay in detecting an intrusion detection system.

3) EFFICIENCY OF CLASSIFICATION ALGORITHMS IN REAL-TIME

From the literature review, it can be observed that most of the classification methods are unable to identify and classify traffic in real-time. However, these methods can classify and identify the traffic in a short period of time after the traffic is generated. The real-time classification plays an important role for ensuring network security and improving QoS. As discussed above, traffic classification needs different procedures such as extraction and selection features, train data, validation, etc. Therefore, the real-time classification of traffic remains a challenging issue in traffic network. The researchers and practitioners need to work on lightweight classification techniques to improve the classifier speed in each process of classification algorithms.

4) CLASSIFICATION FROM UNKNOW TRAFFIC

We observe from the literature review that most of the studies do not obtain the classification from unknown traffic and are

focused on the known traffic. That causes the classification results not accurate in some conditions. For example, if the traffic category does not appear accurately in the training data, the algorithm might classify the traffic as known traffic. Also, note that some methods classify traffic into a known category, but it is very difficult for them to reclassify unknown traffic in different categories. Therefore, the reclassification of unknown traffic remains a challenging issue in existing research that needs to be studied in depth to overcome this issue.

5) NETWORK FUNCTIONALITY FOR TRAFFIC CLASSIFICATION

The network functionality plays an important role in selecting the best classifier. Various network function such as NATing and tunneling influence the performance of classifier [161]. Therefore, the researchers and practitioners must consider these network functions when designing a classifier which may resulted in a better performance of traffic classification algorithm.

6) REAL-TIME APPLICATION

Various ML techniques such as Bayesian Network and Decision Tree have been used for traffic classification. From literature review, we observed that a few researchers employed DPI for P2P applications of traffic classifications [11], [68]. However, security and privacy remains a critical issue while deploying DPI.

Researchers should consider various factors such as traffic speed and big data when designing ML based traffic classification algorithm. These factors could affect the performance of traffic classifier in real-time applications. Moreover, model training and unknown traffic are considered as other limitations for real-time implementation of traffic classification algorithm. Also, parameter selection and its tuning based on network features and characteristics also pose certain challenges and limitations when employing on real environment. Several Big Tech companies are trying to combine ML methods with network functions to overcome these limitations [162].

C. RECOMMENDATIONS FOR ML FOR TRAFFIC CLASSIFICATION

This section discusses the recommendations that researchers needs to consider when employing ML methods for traffic classification. It investigates how importance these recommendations are to researchers and practitioners in order to effectively apply them for obtain a better Internet traffic classification. We outline several recommendations that could be used to enhance the traffic classification framework.

1) GENERALIZE MODEL

The generalized ML model should be tested on datasets collected from different network environments in order to demonstrate the effectiveness of the ML model. For instance, if an ML algorithm is considered a generalized model, it may indicate that the hidden data are present at low variance.

2) ANONYMOUS TRAFFIC DETECTION

The traffic classification aims to identify the traffic characteristic, type, and network application name. These features are constantly emerging in the network environment. Therefore, detecting rapid changes in the network and identifying attacks could be beneficial for reducing the chances of misclassification. In this context, traffic classification model uncertainty need to assess based on the traffic types and features, and a model must have a capability to detect anomalies. The unsupervised ML techniques should be able to detect unknown traffic without requiring any prior knowledge or information of traffic characteristics within the network.

3) ROBUSTNESS OF TRAFFIC CLASSIFICATION

In order to make the classifier more robust for detecting anomaly detection or identifying traffic obfuscation plays an important role to reduce the chances of misclassification. In this regard, the unsupervised learning could be used to detect different unknown traffic classes. Therefore, the traffic classifier is necessary to test against anomaly and intrusion detection system along with different obfuscation approaches. Consequently, we could detect various kinds of attacks or traffic mutations in order to obtain a better classifier.

4) RIGOROUS ANALYSIS

Developed traffic classification models should be comprehensively analyzed using different tools in order to evaluate their effectiveness and efficiency. This can be accomplished using performance and standard metrics and compare its implementations on various Internet traffic such as encryption, decryption along with the multichannel application flows of different length.

5) UPDATED MODEL

Update model of traffic classification techniques evolving rapidly in network, such as traffic classification in IoT network by considering different kinds of IoT devices. It requires to train the model based on latest trends of traffic types. Developing ML techniques should be employed that uses online training for updating traffic classifiers. In this context, reinforcement learning could be used to update training model by relying on the feedback from the users in terms of QoS, security and privacy issues, and number of false alarms of network security.

6) DATA COLLECTION

Data collection plays an important role for assessing the performance of ML based classifier. Training the ML model in terms of representative data that could be used to obtain the useful information or pattern and helps to classify the hidden data with higher accuracy.

Data should be accurately labelled and ensure its collection from different network edges and points. Also, collection of data from a variety of sources such as devices, applications,

etc. could be useful for implementing ML model. Furthermore, the availability of open-access or public data could be used to assess the effectiveness of the developed ML algorithms. Handling big data requires large amount of data to process. The tool required to process the ML technique using big data could be beneficial for obtaining high speed traffic, but it requires large storage techniques to process the ML method with big data.

VIII. CONCLUSION

The research interest on traffic classification has been gaining popularity among researchers from Communication and Networking backgrounds over the last couple of years. Through this technique, network operators could monitor the performance of the traffic classification such as service identifications, network designing, and perform the optimization to classify and identify traffic. It has produced robust and novel results and attained a better accuracy when it applies to different behaviors of Internet applications. Current investigation on emerging trends of traffic classification methods is necessary for researchers, practitioners, and Internet service providers who can monitor the performance of the entire classification network. This paper gave a thorough review of the network traffic classification techniques, traffic datasets, and ML-based methods for traffic classification. We first introduced the traffic classification procedures, in which we thoroughly reviewed the datasets and discussed the extraction and feature selection methods that are widely used in traffic classification. Then, we further presented the criteria for traffic classification, which can be used to assess the effectiveness of classification algorithms. We thoroughly discussed the recent state-of-the-art techniques for traffic classification in terms of four categories such as port-based classification, payload-based classification, statistical-based classification, and behavior-based classification. Then, we discussed the ML methods for traffic classification, which is followed by a thorough discussion of traffic obfuscation techniques. Finally, key findings and open research challenges are identified along with recommendations for future research directions in traffic classification. In short, this survey is well developed to cover traffic classification techniques. It fills the literature gaps of existing surveys and incorporates the recent trends and approaches in traffic classifications.

REFERENCES

[1] *WAN and Application Optimization Solution Guide*, Cisco Validated Design, USA, 2008.

[2] M. Shafiq, Z. Tian, A. K. Bashir, A. Jolfaei, and X. Yu, "Data mining and machine learning methods for sustainable smart cities traffic classification: A survey," *Sustain. Cities Soc.*, vol. 60, Sep. 2020, Art. no. 102177.

[3] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 56–76, 4th Quart., 2008.

[4] (Aug. 14, 2007). *Snort—The de Facto Standard for Intrusion Detection/Prevention*. [Online]. Available: http://www.snort.org

[5] V. Paxson, "Bro: A system for detecting network intruders in real-time," *Comput. Netw.*, vol. 31, nos. 23–24, pp. 2435–2463, Dec. 1999.

[6] L. Peng, H. Zhang, Y. Chen, and B. Yang, "Imbalanced traffic identification using an imbalanced data gravitation-based classification model," *Comput. Commun.*, vol. 102, pp. 177–189, Apr. 2017.

[7] A. Saeed and M. Kolberg, "Towards optimizing WLANs power saving: Novel context-aware network traffic classification based on a machine learning approach," *IEEE Access*, vol. 7, pp. 3122–3135, 2019.

[8] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Semi-supervised network traffic classification," in *Proc. ACM Int. Conf. Meas. Modeling Comput. Syst. (SIGMETRICS) Perform. Eval. Rev.*, 2007, pp. 369–370.

[9] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," in *Proc. 4th ACM SIGCOMM Conf. Internet Meas. (IMC)*, Taormina, Italy, 2004, pp. 135–148.

[10] H. N. Datir and P. M. Jawandhiya, "Survey on hybrid data mining algorithms for intrusion detection system," in *Data Management, Analytics and Innovation*. Singapore: Springer, 2019.

[11] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy, "Transport layer identification of P2P traffic," in *Proc. 4th ACM SIGCOMM Conf. Internet Meas. (IMC)*, 2004, pp. 121–134.

[12] G. Ren, Y. Cao, S. Wen, T. Huang, and Z. Zeng, "A modified Elman neural network with a new learning rate scheme," *Neurocomputing*, vol. 286, pp. 11–18, Apr. 2018.

[13] M. R. Shahid, G. Blanc, Z. Zhang, and H. Debar, "IoT devices recognition through network traffic analysis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5187–5192.

[14] M. Finsterbusch, C. Richter, E. Rocha, J.-A. Müller, and K. Hanssgen, "A survey of payload-based traffic classification approaches," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 1135–1156, 2nd Quart., 2014.

[15] M. Piskozub, R. Spolaor, and I. Martinovic, "MalAlert: Detecting malware in large-scale network traffic using statistical features," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 46, no. 3, pp. 151–154, Jan. 2019.

[16] R. Chapaneri and S. Shah, "A comprehensive survey of machine learning-based network intrusion detection," in *Smart Intelligent Computing and Applications, Smart Innovation, Systems and Technologies*, vol. 104. Singapore: Springer, 2019, doi: 10.1007/978-981-13-1921-1_35.

[17] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur.*, vol. 28, nos. 1–2, pp. 18–28, Feb. 2009.

[18] A. Callado, C. Kamienski, G. Szabo, B. Peter Gero, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 37–52, 3rd Quart., 2009.

[19] M. Bhatia and M. K. Rai, "Identifying P2P traffic: A survey," *Peer-Peer Netw. Appl.*, vol. 10, no. 5, pp. 1182–1203, Sep. 2017.

[20] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[21] A. Dainotti, A. Pescapé, and K. C. Claffy, "Issues and future directions in traffic classification," *IEEE Netw.*, vol. 26, no. 1, pp. 35–40, Jan./Feb. 2012.

[22] M. A. Alsheikh, S. Lin, D. Niyato, and H. P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1996–2018, 4th Quart., 2014.

[23] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *Int. J. Netw. Manage.*, vol. 25, no. 5, pp. 355–374, 2015.

[24] J. V. Gomes, P. R. M. Inácio, M. Pereira, M. M. Freire, and P. P. Monteiro, "Detection and classification of peer-to-peer traffic: A survey," *ACM Comput. Surv.*, vol. 45, no. 3, pp. 1–40, Jun. 2013.

[25] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, "Towards the deployment of machine learning solutions in network traffic classification: A systematic survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1988–2014, 2nd Quart., 2019.

[26] H. Tahaei, F. Afifi, A. Asemi, F. Zaki, and N. B. Anuar, "The rise of traffic classification in IoT networks: A survey," *J. Netw. Comput. Appl.*, vol. 154, Mar. 2020, Art. no. 102538.

[27] A. Akande, P. Cabral, P. Gomes, and S. Casteleyn, "The Lisbon ranking for smart sustainable cities in Europe," *Sustain. Cities Soc.*, vol. 44, pp. 475–487, Jan. 2019.

[28] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, 1st Quart., 2014.

[29] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Towards generating real-life datasets for network intrusion detection," *IJ Netw. Secur.*, vol. 17, no. 6, pp. 683–701, 2015.

[30] F. Gringoli, L. Salgarelli, M. Dusi, N. Cascarano, F. Risso, and K. C. Claffy, "GT: Picking up the truth from the ground for internet traffic," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 5, pp. 12–18, Oct. 2009.
[31] J. McHugh, "Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 262–294, Nov. 2000.
[32] V. Jacobson, C. Leres, and S. McCanne, "TCPDUMP manual page," Lawrence Berkeley Nat. Lab., Univ. California, Berkeley, CA, USA, Tech. Rep., 2020.
[33] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Comput. Secur.*, vol. 31, no. 3, pp. 357–374, 2012.
[34] (2017). *UNB-ISCX Intrusion Detection Evaluation Dataset (CICIDS2017)*. [Online]. Available: https: //www.unb.ca/cic/datasets/ids-2017.html
[35] KDD99. *KDD Cup 1999 Data*. Accessed: Mar. 25, 2021. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[36] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–6.
[37] S. E. Bibri, "The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability," *Sustain. Cities Soc.*, vol. 38, pp. 230–253, Apr. 2018.
[38] AWID. (2014). *AWID Dataset—Wireless Security Datasets Project*. [Online]. Available: http://icsd web.aegean.gr/awid/
[39] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.
[40] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.
[41] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, D. Breitenbacher, A. Shabtai, and Y. Elovici. (2018). *N-BaIoT Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT
[42] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 12–22, Jul. 2018.
[43] CTU University. (2016). *The Stratosphere IPS Project Dataset*. [Online]. Available: https://stratosphere ips.org/category/dataset.html
[44] WAND Research Group. *Wits: Waikato Internet Traffic Storage*. Accessed: Sep. 15, 2021. [Online]. Available: http://wand.net.nz/wits/index.php
[45] L. Peng, B. Yang, Y. Chen, and Z. Chen, "Effectiveness of statistical features for early stage internet traffic identification," *Int. J. Parallel Program.*, vol. 44, no. 1, pp. 181–197, Feb. 2016.
[46] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 184–208, 1st Quart., 2016.
[47] (2017). *CICIDS2017*. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html
[48] D. Kurniabudi, D. Stiawan, M. Y. Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 dataset feature analysis with information gain for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020.
[49] Y. Dhote, S. Agrawal, and A. J. Deen, "A survey on feature selection techniques for internet traffic classification," in *Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Dec. 2015, pp. 1375–1380.
[50] M. Wasikowski and X.-W. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1388–1400, Oct. 2010.
[51] Y. S. Lim, H. C. Kim, J. Jeong, C.-K. Kim, T. T. Kwon, and Y. Choi, "Internet traffic classification demystified: On the sources of the discriminative power," in *Proc. 6th Int. Conf.*, 2010, p. 9.
[52] A. Moore, D. Zuev, and M. Crogan, "Discriminators for use in flow-based classification," Dept. Comput. Sci., Queen Mary Univ. London, London, U.K., Tech. Rep. RR-05-13, pp. 6–13, 2013.
[53] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 2, pp. 23–26, 2006.
[54] S. Dey, Q. Ye, and S. Sampalli, "A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks," *Inf. Fusion*, vol. 49, pp. 205–215, Sep. 2019.
[55] S. Ding, H. Zhu, W. Jia, and C. Su, "A survey on feature extraction for pattern recognition," *Artif. Intell. Rev.*, vol. 37, no. 3, pp. 169–180, 2012.
[56] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, Dec. 2015.
[57] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, "Feature selection for optimizing traffic classification," *Comput. Commun.*, vol. 35, no. 12, pp. 1457–1471, 2012.
[58] Z. Chen, L. Peng, S. Zhao, L. Zhang, and S. Jing, "Feature selection toward optimizing internet traffic behavior identification," in *Proc. Int. Conf. Algorithms Archit. Parallel Process.* Cham, Switzerland: Springer, 2014, pp. 631–644.
[59] J. Yan, "A survey of traffic classification validation and ground truth collection," in *Proc. 8th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jun. 2018, pp. 255–259.
[60] J. Erman, A. Mahanti, and M. Arlitt, "Byte me: A case for byte accuracy in traffic classification," in *Proc. 3rd Annu. ACM Workshop Mining Netw. Data (MineNet)*, 2007, pp. 35–38.
[61] Z. Cao, G. Xiong, Y. Zhao, Z. Li, and L. Guo, "A survey on encrypted traffic classification," in *Proc. Int. Conf. Appl. Techn. Inf. Secur.*, in Communications in Computer and Information Science, vol. 490, 2014, pp. 73–81.
[62] Q. Ma, W. Huang, Y. Jin, and J. Mao, "Encrypted traffic classification based on traffic reconstruction," in *Proc. 4th Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2021, pp. 572–576.
[63] J. Hurley, E. Garcia-Palacios, and S. Sezer, "Classification of P2P and HTTP using specific protocol characteristics," in *Proc. Meeting Eur. Netw. Univ. Companies Inf. Commun. Eng.*, 2009, pp. 31–40.
[64] M. Mohammadi, B. Raahemi, A. Akbari, H. Moeinzadeh, and B. Nasersharif, "Genetic-based minimum classification error mapping for accurate identifying peer-to-peer applications in the internet traffic," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 6417–6423, Jun. 2011.
[65] H. Schulze and K. Mochalski, "Internet study 2008/2009," *Ipoque Rep.*, vol. 37, pp. 351–362, 2009.
[66] G. De La Torre Parra, P. Rad, and K.-K.-R. Choo, "Implementation of deep packet inspection in smart grids and industrial Internet of Things: Challenges and opportunities," *J. Netw. Comput. Appl.*, vol. 135, pp. 32–46, Jun. 2019.
[67] A. W. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in *Proc. Int. Workshop Passive Act. Netw. Meas.*, 2005, pp. 41–45.
[68] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of P2P traffic using application signatures," in *Proc. 13th Conf. World Wide Web (WWW)*, 2004, pp. 512–521.
[69] Z. Liu, R. Wang, and D. Tang, "Extending labeled mobile network traffic data by three levels traffic identification fusion," *Future Gener. Comput. Syst.*, vol. 88, pp. 453–466, Nov. 2018.
[70] J. J. Yang, Narantuya, and H. Lim, "Bayesian neural network based encrypted traffic classification using initial handshake packets," in *Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.-Supplemental (DSN-S)*, Jun. 2019, pp. 19–20.
[71] N. Al Khater and R. E. Overill, "Network traffic classification techniques and challenges," in *Proc. 10th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Oct. 2015, pp. 43–48.
[72] A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 33, no. 1, pp. 50–60, Jun. 2005.
[73] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu, "Robust network traffic classification," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1257–1270, Aug. 2015.
[74] R. Alshammari and A. Nur Zincir-Heywood, "Identification of VoIP encrypted traffic using a machine learning approach," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 27, no. 1, pp. 77–92, Jan. 2015.
[75] V. A. Muliukha, L. U. Laboshin, A. A. Lukashin, and N. V. Nashivochnikov, "Analysis and classification of encrypted network traffic using machine learning," in *Proc. Int. Conf. Soft Comput. Meas.*, 2020, pp. 194–197.
[76] F. Al-Obaidy, S. Momtahen, M. F. Hossain, and F. Mohammadi, "Encrypted traffic classification based ML for identifying different social media applications," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, May 2019, pp. 1–5.
[77] K. Xu, Z. L. Zhang, and S. Bhattacharyya, "Profiling internet backbone traffic: Behavior models and applications," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 169–180, 2005.

[78] Y. Jin, N. Duffield, P. Haffner, S. Sen, and Z.-L. Zhang, "Inferring applications at the network layer using collective traffic statistics," in *Proc. 22nd Int. Teletraffic Congr. (LTC)*, Sep. 2010, pp. 1–8.

[79] P. Bermolen, M. Mellia, M. Meo, D. Rossi, and S. Valenti, "Abacus: Accurate behavioral classification of P2P-TV traffic," *Comput. Netw.*, vol. 55, no. 6, pp. 1394–1411, Apr. 2011.

[80] J. Kohout, T. Komárek, P. Čech, J. Bodnár, and J. Lokoč, "Learning communication patterns for malware discovery in HTTPs data," *Expert Syst. Appl.*, vol. 101, pp. 129–142, Jul. 2018.

[81] S. Maeda, A. Kanai, S. Tanimoto, T. Hatashima, and K. Ohkubo, "A botnet detection method on SDN using deep learning," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2019, pp. 1–6.

[82] C.-L. Hsieh, N. Weng, and W. Wei, "Scalable many-field packet classification for traffic steering in SDN switches," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 1, pp. 348–361, Mar. 2019.

[83] C. Livadas, R. Walsh, D. E. Lapsley, and W. T. Strayer, "Using machine learning techniques to identify botnet traffic," in *Proc. LCN*, 2006, pp. 967–974.

[84] J. Gao, S. Chai, B. Zhang, and Y. Xia, "Research on network intrusion detection based on incremental extreme learning machine and adaptive principal component analysis," *Energies*, vol. 12, no. 7, p. 1223, Mar. 2019.

[85] M. Shafiq and X. Yu, "Effective packet number for 5G IM WeChat application at early stage traffic classification," *Mobile Inf. Syst.*, vol. 2017, pp. 1–22, Feb. 2017.

[86] A. Moore and D. Zuev, "'Internet traffic classification using Bayesian analysis techniques,'" in *Proc. ACM Int. Conf. Meas. Modeling Comput. Syst. (SIGMETRICS)*, Banff, AB, Canada, 2005, pp. 50–60.

[87] J. Park, H.-R. Tyan, and C.-C. Kuo, "Internet traffic classification for scalable QoS provision," in *Proc. IEEE Int. Conf. Multimedia Expo*, Toronto, ON, Canada, Jul. 2006, pp. 1221–1224.

[88] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 5, pp. 5–16, Oct. 2006.

[89] W. Zhou, L. Dong, L. Bic, M. Zhou, and L. Chen, "Internet traffic classification using feed-forward neural network," in *Proc. Int. Conf. Comput. Problem-Solving (ICCP)*, Oct. 2011, pp. 641–646.

[90] M. Cui, J. Wang, and M. Yue, "Machine learning-based anomaly detection for load forecasting under cyberattacks," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5724–5734, Sep. 2019.

[91] A. Bivens, C. Palagiri, R. Smith, B. Szymanski, and M. Embrechts, "Network based intrusion detection using neural networks," *Intell. Eng. Syst. Through Artif. Neural Netw.*, vol. 12, no. 1, pp. 579–584, 2002.

[92] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 686–728, 1st Quart., 2019.

[93] J. Erman, A. Mahanti, M. Arlitt, and C. Williamson, "Identifying and discriminating between web and peer-to-peer traffic in the network core," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, Banff, AB, Canada, 2007, pp. 883–892.

[94] N. Kant and M. Mahajan, "Time-series outlier detection using enhanced K-means in combination with PSO algorithm," in *Engineering Vibration, Communication and Information Processing*. Singapore: Springer, 2019, pp. 363–373.

[95] P. K. Jain and R. Pamula, "Two-step anomaly detection approach using clustering algorithm," in *Proc. Int. Conf. Adv. Comput. Netw. Inform.*, 2019, pp. 513–520.

[96] J. Gauci, E. Conti, Y. Liang, K. Virochsiri, Y. He, Z. Kaden, V. Narayanan, X. Ye, Z. Chen, and S. Fujimoto, "Horizon: Facebook's open source applied reinforcement learning platform," 2018, *arXiv:1811.00260*.

[97] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "QT-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," 2018, *arXiv:1806.10293*.

[98] C. Wagner, J. François, and T. Engel, "Machine learning approach for IP-flow record anomaly detection," in *Proc. Int. Conf. Res. Netw.*, 2011, pp. 28–39.

[99] A.-M. Yang, S.-Y. Jiang, and H. Deng, "A P2P network traffic classification method using SVM," in *Proc. 9th Int. Conf. Young Comput. Sci.*, Nov. 2008, pp. 398–403.

[100] M. Gyanchandani, J. Rana, and R. Yadav, "Taxonomy of anomaly based intrusion detection system: A review," *Int. J. Sci. Res. Publications*, vol. 2, no. 12, pp. 1–13, 2012.

[101] V. E. Mirzakhanov, "Value of fuzzy logic for data mining and machine learning: A case study," *Expert Syst. Appl.*, vol. 162, Dec. 2020, Art. no. 113781.

[102] S. Raja and S. Ramaiah, "An efficient fuzzy-based hybrid system to cloud intrusion detection," *Int. J. Fuzzy Syst.*, vol. 19, no. 1, pp. 62–77, 2016.

[103] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu, C. Wang, and Y. Liu, "A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 393–430, 1st Quart., 2019.

[104] N. McLaughlin, J. M. del Rincon, B. Kang, S. Yerima, P. Miller, S. Sezer, Y. Safaei, E. Trickel, Z. Zhao, A. Doupé, and G. J. Ahn, "Deep Android malware detection," in *Proc. 7th ACM Conf. Data Appl. Secur. Privacy*, 2017, pp. 301–308.

[105] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, 2017, pp. 712–717.

[106] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 43–48.

[107] S. Rezaei and X. Liu, "How to achieve high classification accuracy with just a few labels: A semi-supervised approach using sampled packets," 2018, *arXiv:1812.09761*.

[108] X. Wang, Y. Liu, and W. Su, "Real-time classification method of network traffic based on parallelized CNN," in *Proc. IEEE Int. Conf. Power, Intell. Comput. Syst. (ICPICS)*, Jul. 2019, pp. 92–97.

[109] H. Zhou, Y. Wang, and M. Ye, "A method of CNN traffic classification based on sppnet," in *Proc. 14th Int. Conf. Comput. Intell. Secur. (CIS)*, Nov. 2018, pp. 390–394.

[110] O. Salman, I. H. Elhajj, A. Chehab, and A. Kayssi, "A multi-level internet traffic classifier using deep learning," in *Proc. 9th Int. Conf. Netw. Future (NOF)*, Nov. 2018, pp. 68–75.

[111] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton, "Peek-a-boo, I still see you: Why efficient traffic analysis countermeasures fail," in *Proc. IEEE Symp. Secur. Privacy*, May 2012, pp. 332–346.

[112] *Pluggable-Transports/Obfsproxy—Pluggable Transport for Obfuscated Traffic*. Accessed: Sep. 18, 2021. [Online]. Available: https://gitweb.torproject.org/pluggable-transports/obfsproxy.git/tree/doc/obfs2/obfs2-protocol-spec.txt

[113] *Flash Proxies*. Accessed: Sep. 18, 2021. [Online]. Available: https://crypto.stanford.edu/flashproxy/

[114] A. Houmansadr, T. Riedl, S. N. Borisov, and A. Inger, "I want my voice to be heard: IP over voice-over-IP for unobservable censorship circumvention," in *Proc. NDSS*, 2013, pp. 1–17.

[115] *Doc/Meek—Tor Bug Tracker & Wiki*. Accessed: Oct. 25, 2021. [Online]. Available: https://trac.torproject.org/ projects/tor/wiki/doc/meek

[116] J. Gardiner and S. Nagaraja, "Blindspot: Indistinguishable anonymous communications," 2014, *arXiv:1408.0784*.

[117] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton, "Protocol misidentification made easy with format-transforming encryption," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2013, pp. 61–72.

[118] J. Lv, C. Zhu, S. Tang, and C. Yang, "Deepflow: Hiding anonymous communication traffic in P2P streaming networks," *Wuhan Univ. J. Nat. Sci.*, vol. 19, no. 5, pp. 417–425, 2014.

[119] S. Li, M. Schliep, and N. Hopper, "Facet: Streaming over videoconferencing for censorship circumvention," in *Proc. 13th Workshop Privacy Electron. Soc.*, Nov. 2014, pp. 163–172.

[120] *Protocol Obfuscation—Emule Wiki*. Accessed: Sep. 18, 2021. [Online]. Available: http://wiki.emule-web.de/ Protocol obfuscation

[121] F. Li, A. Razaghpanah, A. M. Kakhki, A. A. Niaki, D. Choffnes, P. Gill, and A. Mislove, "Liberate,(n): A library for exposing (trafficclassification) rules and avoiding them efficiently," in *Proc. Internet Meas. Conf.*, 2017, pp. 128–141.

[122] W. Moore, H. Tan, M. Sher, and M. Maloof, "Multiclass traffic morphing for encrypted VoIP communication," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, 2015, pp. 65–85.

[123] R. McPherson, A. Houmansadr, and V. Shmatikov, "CovertCast: Using live streaming to evade internet censorship," in *Proc. Privacy Enhancing Technol.*, 2016, pp. 212–225.

[124] X. Cai, R. Nithyanand, T. Wang, R. Johnson, and I. Goldberg, "A systematic approach to developing and evaluating website fingerprinting defenses," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 227–238.
[125] D. Barradas, N. Santos, and L. Rodrigues, "DeltaShaper: Enabling unobservable censorship-resistant TCP tunneling over videoconferencing streams," in *Proc. Privacy Enhancing Technol.*, 2017, pp. 5–22.
[126] K. Kohls, T. Holz, D. Kolossa, and C. Pöpper, "SkypeLine: Robust hidden data transmission for VoIP," in *Proc. 11th ACM Asia Conf. Comput. Commun. Secur.*, May 2016, pp. 877–888.
[127] D. Heckerman, "A tutorial on learning with Bayesian networks," in *Innovations in Bayesian Networks*. Berlin, Germany: Springer, 2008.
[128] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 223–239, Jan. 2007.
[129] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.
[130] M. Shafiq, X. Yu, and A. A. Laghari, "WeChat text messages service flow traffic classification using machine learning technique," in *Proc. 6th Int. Conf. IT Converg. Secur. (ICITCS)*, Sep. 2016, pp. 1–5.
[131] J. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, Mar. 1986.
[132] S. L. Salzberg, "C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, 1994.
[133] M. Shafiq, X. Yu, A. A. Laghari, and D. Wang, "Effective feature selection for 5G IM applications traffic classification," *Mobile Inf. Syst.*, vol. 2017, pp. 1–12, May 2017.
[134] V. S. Dave and K. Dutta, "Neural network-based models for software effort estimation: A review," *Artif. Intell. Rev.*, vol. 42, no. 2, pp. 295–307, 2014.
[135] N. Izeboudjen, C. Larbes, and A. Farah, "A new classification approach for neural networks hardware: From standards chips to embedded systems on chip," *Artif. Intell. Rev.*, vol. 41, no. 4, pp. 491–537, 2014.
[136] R. Sun, B. Yang, L. Peng, Z. Chen, L. Zhang, and S. Jing, "Traffic classification using probabilistic neural networks," in *Proc. 6th Int. Conf. Natural Comput.*, vol. 4, 2010, pp. 1914–1919.
[137] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Found. Trends Mach. Learn.*, vol. 11, nos. 3–4, pp. 219–354, 2018.
[138] L.-L. Wang, H. Y. T. Ngan, and N. H. C. Yung, "Automatic incident classification for large-scale traffic data by adaptive boosting SVM," *Inf. Sci.*, vol. 467, pp. 59–73, Oct. 2018.
[139] J. Nash, "Non-cooperative games," *Ann. Math.*, vol. 54, no. 2, pp. 286–295, 1951.
[140] S. Rezaei and X. Liu, "Deep learning for encrypted traffic classification: An overview," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 76–81, Dec. 2019.
[141] Z. Wang, "The applications of deep learning on traffic identification," *BlackHat USA*, vol. 24, no. 11, pp. 1–10, 2015.
[142] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network traffic classifier with convolutional and recurrent neural networks for Internet of Things," *IEEE Access*, vol. 5, pp. 18042–18050, 2017.
[143] T. Iwai and A. Nakao, "Adaptive mobile application identification through in-network machine learning," in *Proc. 18th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Oct. 2016, pp. 1–6.
[144] K. P. Dyer, S. E. Coull, T. Ristenpart, and T. Shrimpton, "Peek-a-boo, I still see you: Why efficient traffic analysis countermeasures fail," in *Proc. IEEE Symp. Secur. Privacy*, May 2012, pp. 332–346.
[145] X. Cai, R. Nithyanand, T. Wang, R. Johnson, and I. Goldberg, "A systematic approach to developing and evaluating website fingerprinting defenses," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 227–238.
[146] T. Tabatabaei, M. Adel, F. Karray, and M. Kamel, "Machine learning-based classification of encrypted internet traffic," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.* Berlin, Germany: Springer, 2012, pp. 578–592.
[147] Y. Liu, J. Chen, P. Chang, and X. Yun, "A novel algorithm for encrypted traffic classification based on sliding window of flow's first n packets," in *Proc. 2nd IEEE Int. Conf. Comput. Intell. Appl. (ICCIA)*, Sep. 2017, pp. 463–470.
[148] A. Parchekani, S. Nouri, V. Shah-Mansouri, and S. P. Shariatpanahi, "Classification of traffic using neural networks by rejecting: A novel approach in classifying VPN traffic," 2020, *arXiv:2001.03665*.
[149] Q. Sun, D. Simon, Y. Wang, W. Russell, V. Padmanabhan, and L. Qiu, "Statistical identification of encrypted web browsing traffic," in *Proc. IEEE Symp. Secur. Privacy*, May 2002, pp. 19–30.
[150] M. Gruteser and D. Grunwald, "Enhancing location privacy in wireless LAN through disposable interface identifiers: A quantitative analysis," *Mobile Netw. Appl.*, vol. 10, no. 3, pp. 315–325, 2005.
[151] F. Zhang, W. He, and X. Liu, "Defending against traffic analysis in wireless networks through traffic reshaping," in *Proc. 31st Int. Conf. Distrib. Comput. Syst.*, Jun. 2011, pp. 593–602.
[152] H. Xiao, B. Xiao, and Y. Huang, "Implementation of covert communication based on steganography," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Aug. 2008, pp. 1512–1515.
[153] L. Invernizzi, C. Kruegel, and G. Vigna, "Message in a bottle: Sailing past censorship," in *Proc. 29th Annu. Comput. Secur. Appl. Conf.*, Dec. 2013, pp. 39–48.
[154] C. Wright, S. Coull, and F. Monrose, "Traffic morphing: An efficient defense against statistical traffic analysis," in *Proc. NDSS*, vol. 9, 2009, pp. 1–14.
[155] L. Bernaille and R. Teixeira, "Early recognition of encrypted applications," in *Proc. Int. Conf. Passive Active Netw. Meas.* Berlin, Germany: Springer, 2007, pp. 165–175.
[156] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Robust smartphone app identification via encrypted network traffic analysis," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 63–78, Jan. 2018.
[157] L. Wang, K. P. Dyer, A. Akella, T. Ristenpart, and T. Shrimpton, "Seeing through network-protocol obfuscation," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 57–69.
[158] H. M. Moghaddam, B. Li, M. Derakhshani, and I. Goldberg, "SkypeMorph: Protocol obfuscation for tor bridges," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2012, pp. 97–108.
[159] Z. Weinberg, J. Wang, V. Yegneswaran, L. Briesemeister, S. Cheung, F. Wang, and D. Boneh, "StegoTorus: A camouflage proxy for the tor anonymity system," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, 2012, pp. 109–120.
[160] J. Laufs, H. Borrion, and B. Bradford, "Security and the smart city: A systematic review," *Sustain. Cities Soc.*, vol. 55, Apr. 2020, Art. no. 102023.
[161] M. Crotti, F. Gringoli, and L. Salgarelli, "Impact of asymmetric routing on statistical traffic classification," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2009, pp. 1–8.
[162] O. Salman, I. Elhajj, A. Kayssi, and C. Ali, "A review on machine learning–based approaches for internet traffic classification," *Ann. Telecommun.*, vol. 75, pp. 673–710, Jun. 2020.

MUHAMMAD SAMEER SHEIKH received the Ph.D. degree in communication and information systems from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2017. He was a Postdoctoral Research Fellow with Jiangsu University, from 2018 to 2020. In 2020, he joined as an Associate Professor at Guangdong Ocean University, China. He has authored more than 20 articles published in international journals and international conference proceeding. His research interests include intelligent transportation systems (ITSs), traffic information engineering, traffic classification, traffic control and operation, traffic networks, connected and automated vehicles, and the IoT technologies. He is a member of the Technical Committee of International Conferences and a reviewer of various reputed journals.

YINQIAO PENG received the Ph.D. degree from Central South University, in 2013. He is currently working as an Associate Professor with the Department of Electronics and Information Engineering, Guangdong Ocean University, China. His research interests include digital signal processing, the IoT technology, and data sciences.