A Complete Review on the Application of Statistical Methods for Evaluating Internet Traffic Usage

Vanice Canuto Cunha, Arturo Zavala Zavala, Damien Magoni, Pedro R M Inacio, Mario M Freire

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 November 2022, accepted 2 December 2022, date of publication 6 December 2022,
date of current version 13 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227073

TOPICAL REVIEW

A Complete Review on the Application of Statistical Methods for Evaluating Internet Traffic Usage

VANICE CANUTO CUNHA[1,2], ARTURO ZAVALA ZAVALA[3], DAMIEN MAGONI[4], (Senior Member, IEEE), PEDRO R. M. INÁCIO[2], (Senior Member, IEEE), AND MÁRIO M. FREIRE[2], (Member, IEEE)

[1]Instituto de Computação (IC), Universidade Federal de Mato Grosso, Cuiabá 78060-900, Brazil
[2]Instituto de Telecomunicações, Universidade da Beira Interior, 6201-001 Covilhã, Portugal
[3]Faculdade de Economia (FE), Universidade Federal de Mato Grosso, Cuiabá 78060-900, Brazil
[4]LaBRI-CNRS, Université de Bordeaux, 33405 Talence, France

Corresponding author: Mário M. Freire (mario@di.ubi.pt)

This work was supported in part by the CAPES (Brazilian Federal Agency for Support and Evaluation of Graduate Education) within the Ministry of Education of Brazil under a scholarship supported by the International Cooperation Program CAPES/Comitê Francês de Avaliação da Cooperação Universitária com o Brasil (COFECUB) at the University of Beira Interior under Project 9090-13-4/2013; in part by the Fundação para a Ciência e a Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through national funds and, when applicable, co-funded by European Union (EU) funds under Project UIDB/50008/2020; in part by the FCT/Programa Operacional Competitividade e Internacionalização (COMPETE)/Fundo Europeu de Desenvolvimento Regional (FEDER) through the Project Towards the assurance of SECURity by dESIGN of the Internet of Things (SECURIoTESIGN) under Grant POCI-01-0145-FEDER-030657; and in part by the Operation Centro-01-0145-FEDER-000019-C4-Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica–Programas Integrados de Investigação Científica e Desenvolvimento Tecnológico (IC&DT).

ABSTRACT Internet traffic classification aims to identify the kind of Internet traffic. With the rise of traffic encryption and multi-layer data encapsulation, some classic classification methods have lost their strength. In an attempt to increase classification performance, Machine Learning (ML) strategies have gained the scientific community interest and have shown themselves promising in the future of traffic classification, mainly in the recognition of encrypted traffic. However, some of these methods have a high computational resource consumption, which make them unfeasible for classification of large traffic flows or in real-time. Methods using statistical analysis have been used to classify real-time traffic or large traffic flows, where the main objective is to find statistical differences among flows or find a pattern in traffic characteristics through statistical properties that allow traffic classification. The purpose of this work is to address statistical methods to classify Internet traffic that were little or unexplored in the literature. This work is not generally focused on discussing statistical methodology. It focuses on discussing statistical tools applied to Internet traffic classification Thus, we provide an overview on statistical distances and divergences previously used or with potential to be used in the classification of Internet traffic. Then, we review previous works about Internet traffic classification using statistical methods, namely Euclidean, Bhattacharyya, and Hellinger distances, Jensen-Shannon and Kullback–Leibler (KL) divergences, Support Vector Machines (SVM), Correlation Information (Pearson Correlation), Kolmogorov-Smirnov and Chi-Square tests, and Entropy. We also discuss some open issues and future research directions on Internet traffic classification using statistical methods.

INDEX TERMS Encrypted internet traffic, traffic classification, statistical distances, statistical divergences, statistical methods, support vector machines.

The associate editor coordinating the review of this manuscript and approving it for publication was Hosam El-Ocla.

I. INTRODUCTION

Internet traffic classification may be used to solve several kinds of network issues. Through traffic classification,

Internet Service Providers (ISP), governments, and network administrators can have access to network resource management, advanced network monitoring, network audit, anomaly detection, and device filtering [1].

Classifying traffic by categorizing network traffic according to its appropriate class is vital to many applications such as pricing, Quality of Service (QoS) control, malware/intrusion detection, and resource usage planning [2].

Due to the importance of classification, several approaches were thought with the development of different applications and scenarios. However, communication advances like encryption and port obfuscation added new challenges to network traffic classification [2].

According to Zhao et al. [3], to manage, detect intrusion, monitor the network security and classify the traffic in real time and in a precise way, the traffic classification is essential. Traffic classification determines the class of the data, grouping and relating them according to the category, making it essential as technique to control and secure the network, besides that, it can foresee and identify the user's behavior in the network [4]. The identification in the right way of traffic categories generated by different applications and protocols help the network operators and administrators, besides supplying a high QoS to the users [3].

Peng et al. [5] states that network traffic classification is a way to identify protocol and application type, besides classifying the traffic. It is the most vital step to manage modern networks and improve network services [6]. The increase of efforts is essential to improve the efficiency of classifiers based on applications and protocols when managing computer networks [5].

According to Valenti et al. [7], the identification of network applications and protocols is a process known as traffic classification. In the last two decades, this theme has gained space in research and several studies have proposed techniques and methods to classify traffic [4], [6], [8], [9], [10]. Among the more classical techniques, we can find payload-based techniques, ML-based techniques and port-based techniques.

By looking at the port number which an application or protocol uses to, port-based techniques enable us to classify those protocols and applications, based on the Internet Assigned Number Authority (IANA) [11]. There are many problems on port-based techniques, especially when dynamic port numbers are used on new applications to avoid detection [12]. This problem is widely known by researchers and has already been addressed on other researches [12]. A proposed alternative was to search within the packets for data sets that could be used as signature for a target application traffic [8], [13].

Payload-based techniques are also known as Deep Packet Inspection (DPI) [13]. These techniques are an alternative to the port-based techniques and are especially used in Peer-to-Peer (P2P) applications that use random port numbers to stream applications over the network [13]. One of the characteristics of these techniques is the examination of the packets content, regardless of the port number, to find attributes of network traffic protocols and applications [14], [15]. However, these techniques also have problems [16]. When faced with traffic from encrypted network applications, they are not efficient, having a high consumption of hardware resources to inspect the payload of each application and protocol [12]. Due to this disadvantage, methods that do not require DPI, such as "in the dark classification" have been developed [16].

In the dark classification sorts traffic by using behavioral and statistical patterns [16]. Gomez et al. [16] state that identifying the application without examining its packet is the major advantage of in the dark classification. The flow statistical behavior and transport layer information, such as packet length, packet inter-arrival time, Transmission Control Protocol (TCP)/Internet Protocol (IP) flags, and checksums are used for protocol identification. This approach can use a training set of sample traffic as a mechanism to identify and classify future traffic based on the application flow behavior [17]. Identification is done through traffic flow properties, such as packet size, entropy, and so on [17].

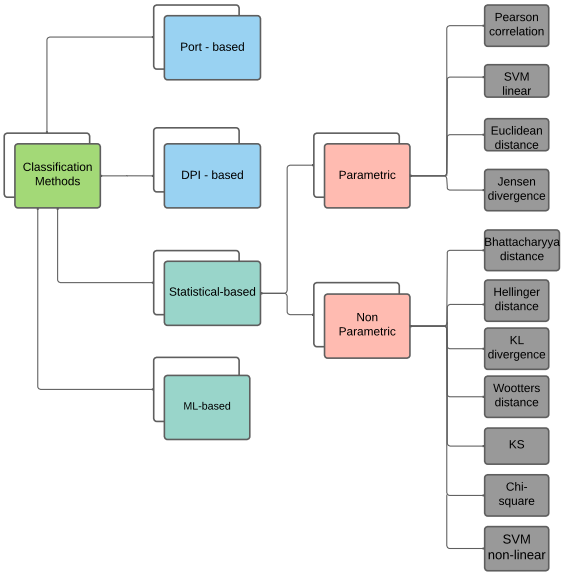

FIGURE 1. Overview of statistical methods for Internet traffic classification.

Different techniques are used to deduce the application protocol and correlate traffic properties, such as Machine Learning (ML) algorithms, sets of heuristics, or statistical measures [7]. For example, according to Liu [17], many researchers use ML to perform statistic-based classification. Statistical classification methods can be divided into two categories: parametric and non-parametric methods [18].

We propose and use the taxonomy of classification methods shown in Figure 1. We address statistical-based methods covering both parametric and non-parametric methods. The category of parametric methods includes linear Support Vector Machines (SVM) [19], Euclidean Distance [20], Pearson correlation [21] and Jensen-Shannon Divergence [22]. The category of non-parametric methods includes non-linear SVM [19], Bhattacharyya Distance [23], Hellinger Distance [24], Kullback-Leibler (KL)

Divergence [25], Wootters Distance [26], and Kolmogorov-Smirnov (KS) [27] and Chi-square [27] tests. Classifiers based in parametric methods have, for each class, a statistical probability distribution. As for the non parametric classifiers, they are used to estimate the statistical probability distribution, or in cases in which the density function is unknown [18].

Many surveys were written about traffic classification. Those surveys summarize the methods and have different focuses as presented on Table 1. The main difference between our research and other review works [2], [3], [8], [10], [16], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], and [6] is that our proposal addresses solutions to solve traffic classification problems by using statistical methods, focusing on distances or divergences. Table 1 presents a comparison with other surveys published in the last ten years that were based on literature from previous decades. For this reason, we emphasized our work on the last ten years, since those papers already considered previous works. Details about ML and Deep Learning can be found at [2], [38], [42] and [6].

In this work, we review the classification of Internet traffic based on statistical methods, including classification methods applied ''in the dark'', observing the main objectives of each survey. It is important to emphasize that we also describe the statistical methods and distances proposed for classification in general, and specific traffic classification found in the literature. Specifying the research method is a crucial step in literature reviews [46]. Our study was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology [47], [48].

According to [48], the PRISMA methodology offers us an evidence-based collection of items, which can be used as a basis for revision work. In addition, PRISMA provides us with a flowchart that allows us to visualize the search strategies and eligibility of the articles. The PRISMA flowchart describes the information cycle used in the different review phases. In order to present and detail our selection process, the flowchart was prepared as shown in Figure 2. The flowchart has 3 phases: identification, screening and included. Through the flowchart, we mapped the number of articles identified, included and excluded, and the exclusions reasons.

The reviewed articles in this paper were chosen from the almost 507 most relevant articles found on a search in IEEE Xplore, Elsevier, ACM Digital Library, Google Scholar and Scopus with the keywords: Internet Traffic Classification, Traffic Classification, Traffic Identification, Encrypted Network Traffic, Network Monitoring, and Statistical Distributions. We also searched for papers with the keywords: Statistical Methods, Statistical Distances, Statistical Analysis, Parametric, Non-Parametric, in the time period extending from 2011 to 2021. In total, 145 articles were reviewed for this work.

Our inclusion criteria were full papers published in journals, articles written in English, and articles that address

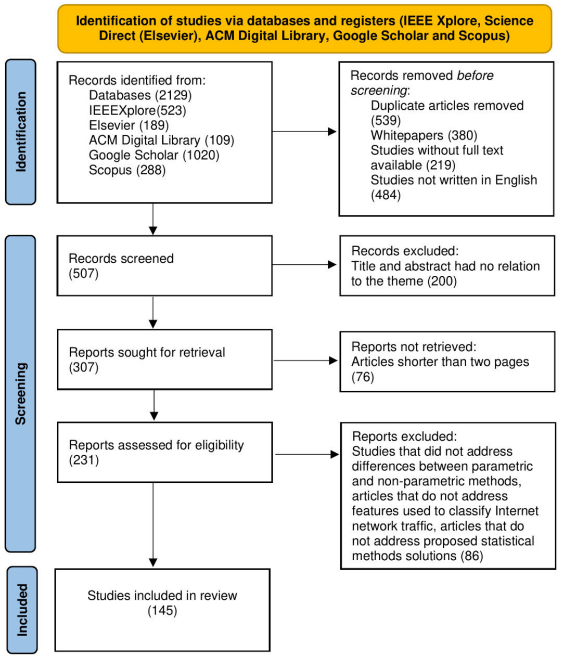

FIGURE 2. The PRISMA flowchart.

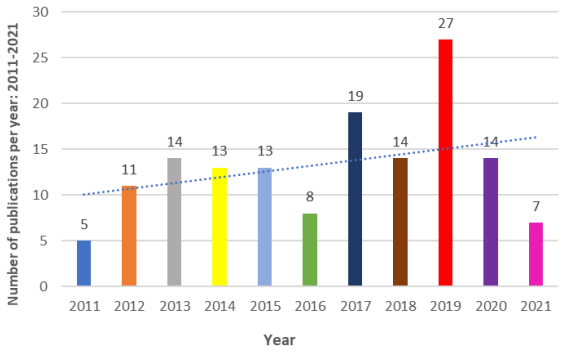

FIGURE 3. Selected studies on statistical methods for internet traffic evaluation (2011 - 2021).

features used to classify Internet network traffic. Our exclusion criteria were duplicate articles, whitepapers, articles shorter than two pages, studies that did not address differences between parametric and non-parametric methods, articles that do not address features used to classify Internet network traffic, articles that do not address proposed statistical methods solutions, articles written in languages other than English, and studies without full text available. After applying the keywords, articles that were not related to the topic in question were excluded by reading the abstract and title. We selected for full reading the articles that could be included after the exclusion and inclusion criteria.

In the initial phase, 2129 articles were identified; of which 1622 were excluded because they were duplicates, whitepapers, studies without full text available, and articles written in languages other than English; 507 were pre-selected. In the screening phase, the abstracts and titles of the articles were read and those unrelated to the topic were excluded, totaling 200 excluded and 307 eligible. Out of the 307 articles,

TABLE 1. Comparison of recent surveys on Internet traffic classification.

Work	Year	Objective	Methods
Dunayts ev *et al.* [28]	2012	Presents the best practices and approaches developed to deal with P2P file sharing traffic, identifying those that may provide long-term benefits for both ISPs and users.	-
Pradhan [29]	2012	Presents a theoretical aspect of SVM, its concepts and applications overview.	SVM
Dainotti *et al.* [30]	2012	Presents reviews and discuss future directions in traffic classification, along with their applicability, reliability, and privacy.	Port/DPI/ML
Gomes *et al.* [16]	2013	Presents the studies on P2P traffic detection and port-based, DPI-based and ML-based classification approaches.	Port/DPI/ML
Se [31]	2013	Presents a survey on several ML techniques for IP traffic classification.	ML
Li *et al.* [32]	2013	Presents studies surveyed about advanced methodologies, such as machine learning datasets, and perspectives.	ML
Finsterbusch *et al.* [8]	2013	Presents a survey focused on performance analysis, technical requirements and accuracy in the DPI rating.	DPI
Dhote *et al. et al.* [40]	2016	Presents a research that addresses feature selection algorithms, focusing on: filter, wrapper, and embedded methods. It also provides an overview of some of the feature selection techniques presented in the literature.	Features - ML
Mehta and Shah [10]	2017	Presents a survey focusing on different types of network classification approaches.	Port/DPI/ML
Yan and Yuan [41]	2018	Examines emerging research on traffic classification techniques in Software-Defined Networks (SDN)	ML
Garrett *et al.* [34]	2018	Researches focused on finding tools and strategies to detect network traffic differentiation	Nearest Neighbor (NN)
Tavara [33]	2019	Presents a summary of parallel algorithmic approaches and parallel tools for SVM implementations focused on efficient approaches and large-scale problem solving.	SVM
Liu and Lang [35]	2019	Classifies and summarize Intrusion Detection Systems (IDSs) based on machine learning focused on solving network security issues.	ML
Rezaei and Liu [2]	2019	Presents a survey on the general structure to rank traffic based on deep learning, as well as the deep learning methods to rank traffic.	Deep Learning (DL)
Nalepa and Kawulok [43]	2019	Presents extensive research on existing methods to select SVM training data from large datasets.	Features - ML
Wang *et al.* [38]	2019	Presents a survey on the general Deep Learning-based mobile traffic classification framework, research approaches to traffic classification focused on mobile encrypted traffic classification in deep learning.	DL
Salman *et al.* [45]	2020	Presents a review of several data representation methods and the different goals of Internet traffic classification.	ML
Alqudah emphet al. [42]	2020	Presents a survey on different machine learning approaches for traffic analysis.	ML
Shen *et al.* [36]	2020	Presents a survey focusing on the systematic approach to optimize feature selection for an efficient classification of encrypted traffic.	Features - ML
Tahaei *et al.* [37]	2020	Provides a review of Internet of Things (IoT) problems and solutions for network traffic classification.	ML
Alam *et al.* [39]	2020	Provides a review focusing on issues related to one-class support vector classifiers.	SVM
Liu and Yu [6]	2021	Presents a survey about encrypted traffic identification focusing on ML.	ML
Zhao *et al.* [3]	2021	Provides a review of network traffic classification methods covering correlation-based, port-based, behavior-based, statistics-based, and payload-based classification.	correlation-based, statistics-based, behavior-based, payload-based, and port-based classification.
Our survey	2022	Presents a study on statistical methods, overviewing statistical distances and divergences for classification of Internet traffic.	Statistics-based, **distance-based**, and **divergence-based** classification.

76 were eventually removed as they were too short, with only two pages. Finally, 231 articles were fully read, of which 86 were excluded for not addressing statistical methods solutions proposed or differences between parametric and non-parametric methods, totaling 145 that met our eligibility criteria and were included in our study. In order to present the results of our selection of articles, following our eligibility criteria, a statistical analytical visualization chart was

generated, as shown in Figure 3 with the number of articles per year.

This work was structured as: review of the classification processes on Section II. Overview of SVM and statistical methods focusing on distances and divergences on Section III. Several methods to classify Internet traffic by using statistical methods on Section IV. Discussion and list of open issues on Section V. Section VI concludes the review.

II. PROCESS OF TRAFFIC CLASSIFICATION: OVERVIEW

A. CLASSIFICATION PROCEDURES

An overview of the traffic classification process was provided in this section as it follows: Internet traffic categorization, data-set, features, classification approach, and validation. Collecting data from a network is a critical point and serves as input to form a pool of network traffic. Extracting and selecting features is an vital process as it can impact the efficiency and effectiveness of classification. The approach chosen for traffic identification is essential to the classification success, as well as ranking performance evaluation criteria [3]. The Figure 4 shows the procedures for classification. As it follows, two topics will be approached: 1) Internet traffic applications, and 2) Dataset.

1) INTERNET TRAFFIC APPLICATIONS

Internet traffic describes the quantity of information or data presented throughout the Web and on different applications, it can be considered a data flow on the Internet [2]. According to [49] Internet traffic is categorized and described according to Table 2. Internet traffic is grouped forming a Dataset.

Internet traffic is divided into ten categories: Administration, Communications, Gaming, Filesharing, Marketplaces, Social Networking, Real-Time Entertainment, Storage, Tunneling, Web Browsing. Each category has a description, that characterizes the associated traffic. The Administration category can be described by services and applications used to administrate the network, such as SNMP, and ICMP protocols. Gaming includes traffic by PC gaming, console, download traffic of consoles, and game updates, like Xbox Live and Playstation traffic. File-sharing includes applications that use distribution protocol models or peer-to-peer, such as Gnutella, eDonkey, Bittorrent, Newsgroups, Ares. Web Browsing includes specific websites, and Web protocols, like WAP browsing, and HTTP.

2) DATASET

Dataset, in classification, has huge importance on evaluating and comparing the performance of different methods. A dataset must contain many diverse samples of each class. A model can fit itself to a specific dataset, doing so, the worry around the probability of the dataset having a deterministic behavior appears. That can happen when a model adjusts itself too much to a specific type of traffic or to a dataset, either because of a lack of interaction to group of users or even for interacting to a small group of users [2]. Usually, for being diversified, traffic is observed on ISPs core, which means that the farther away from the destiny the captured traffic set is, the smaller is the probability of having a deterministic behavior [2].

According to [50], and [42], a dataset is collected and used as an input for training and classification purposes on an ML classifier. On statistical-based methods, statistical resources are allowed to be extracted from flows. These resources are characteristics or properties of flows calculated over many packets [50]. Normally, different features and datasets are used classification-wise.

As stated by [45] a pre-processing phase happens after data collection to extract features that are going to be included into the model. For traffic classification, it is required to evaluate network flow with their main characteristics (packet inter-arrival time and size) with their various statistical values (standard deviation, quartile, min, and max). A set of packets that have the same connection parameters is defined as a flow. Those parameters include port numbers and transport protocol, destination and source IP addresses.

As said in [51], a different way of representing Internet traffic is through time series. They are network flows represented by generating time series of communicated packets/bytes. For each flow three-time series are generated: (1) for bytes channeled through input packets, (2) for bytes channeled through output packets, and (3) for bytes channeled through input and output packets. A short description about feature selection will be present as follows.

B. FEATURE SELECTION

Features are considered in the process of investigating methods and approaches to characterize and classify traffic. Feature selection is a important step in Internet traffic classification. The author [52] sets it as the process of selecting the smallest set of features needed to reach a precise classification. The classification of different application categories occurs when there are some discrepancies in traffic behavior based on the selected features. However, [40] claims that researchers have chosen one or some features from a set of characteristics to classify different traffic flows, basing it only on the qualitative analysis of different features. For analysis purposes, according to [36], it is needed to classify encrypted traffic into numerous flows based on five tuples: User Datagram Protocol (UDP)/TCP, source/destination IP addresses and port numbers. Hereafter the following features will be approached: 1) Packet-Length Based features, 2) Packet-Ordering Based features, and 3) Packet-Timing Based features.

1) PACKET-LENGTH BASED FEATURES

As packet length is a feature related to network packets, according to [36], and [53], its information becomes a commonly used type of resource and it has demonstrated its effectiveness in analyzing traffic that has been encrypted. On packet-length based features, the packet length, the

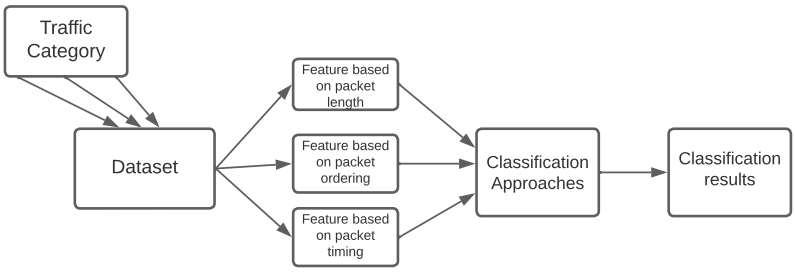

FIGURE 4. Classification procedure.

TABLE 2. Description of traffic categories according to global internet phenomena report: 1H 2014 [49].

Category of Traffic	Explanation	Examples
Administration	Services and applications used for network administration.	SNMP, DNS, NTP, ICMP.
Communications	Protocols, applications and services that allow chat, video and voice communications; information sharing (photos, status, etc) between users.	FaceTime, WhatsApp, Skype, iMessage.
Filesharing	File distribution models or models that use peer-to-peer file sharing.	Newsgroups, eDonkey, BitTorrent, Ares, Gnutella.
Gaming	Application console, game updates and PC gaming download traffic of consoles.	PC games, Playstation 2, Xbox Live, Playstation 3, Nintendo Wii.
Marketplaces	Marketplace apps where purchases of media, apps, books, software, movie, music, and updates are performed by subscribers.	Windows Update, Apple iTunes, Google Android Marketplace.
Real-Time Entertainment	Protocols and applications that provide "on-demand" entertainment that is consumed (viewed or heard) as it arrives.	Buffered or streamed video and audio (RTP, RTSP, MPEG, RTMP, Flash), specific streaming sites, peercasting (Octoshape, PPStream), and services (Spotify,YouTube, Hulu, Netflix).
Social Web	Services and websites that allow interaction such as chats and other types of communication, as well as sharing information between customers and users.	Instagram, Twitter, Facebook, Linkedin.
Storage	It allows transferring through the File Transfer Protocol a massive volume of data, in addition to allowing file hosting, backup and download services.	Dropbox, FTP, zShare, Mozy, Rapidshare, Carbonite.
Tunneling	Services and protocols that mask application identity or allow remote access to network resources.	Remote Desktop, VNC, PC Anywhere, Secure Sockets Layer (SSL), SSH.
Web Browsing	Specific websites and web protocols.	WAP Browsing, HTTP.

cumulative length sequence and the statistics that can be drawn out of the flow, such as minimum, maximum, average, median variance, standard deviation, relative frequency, kurtosis, skew, packet size and variance can be statistical values of packet length.

When obtaining the packet length, in each flow, the first length sequence of the X packets can be used as a key resource. Those X packets can vary a lot length-wise from one website to another, because of their different content and protocol parameters, like those in handshake process, more specifically in the Transport Layer Security (TLS)/SSL. We can use lengths of distinctive packets to distinguish different traffic types.

In a flow, the packet lengths are distributed in intervals that depend on the transport layer and on the MTU (Maximum Transmission Unit). To obtain statistical characteristics of packet length in a flow, the packet length can be aggregated to a fixed number of buckets or bags. To obtain the cumulative length sequence, considering the flow direction, the length of up-link packets can be defined as negative, and as positive for down-link packets. The length of the packets sent are then accumulated to obtain a sequence of the first X cumulative packet lengths. Considering bidirectional flows, we can be define as positive when the packet length is up or down-link. A sequence of cumulative length of the first X packets on a flow seems to be a differentiating feature.

2) PACKET-ORDERING BASED FEATURES

In some cases, the lengths of packets are alike or even the same between different encrypted traffic flows. That makes so alternatives based only on packet length seen less efficient because of the information used. For that, a counter or techniques based on packet counting can be useful [36].

Some packet counts can be considered, such as counting the quantity of up and down-links for each X packets. We can also count the amount of packets before each up-link. Besides that, we can also extract a resource that indicates the number of down-links between two up-link packets.

According to the literature, burst counting can also be useful. An up-link packet burst can be used as example, since down-link packets are exposed to network delays. To do so, the quantity, maximum and average of bursts were considered for each flow.

3) PACKET-TIMING BASED FEATURES

Several information about timestamps of packets can be used to characterize and classify traffic [36], [40]. Inter-Packet Delay is one of the examples. When packets are sent through the network, they receive a timestamp of date and time. The difference between timestamps is defined by the Inter-Packet Delay.

To determine the period of time a transmission is concentrated, the quantity of packets in a time interval is calculated for every series of packets. Timing characteristics generally have limitations, as we most often consider time distributions to be equal, when in reality they are not. The timestamps of packets may experience network fluctuations. This feature can be combined with control packets such as ACKs, CTSs reference points. Table 3 presents a summary of packet-based features.

Even though some features were chosen to classify Internet traffic differently, they do not have the same level of importance. To better understand, each selected feature can receive a weight value that represents its importance. In order to select only important sets of resources, the author in [40] discusses three methods, Wrapper method, Filter method, and Embedded Method, which are briefly described below:

- Wrapper method - Makes use of machine algorithms to rate the performance of different subsets to aid learning. The results are not specific to the ML algorithms used, for this process Genetic Algorithm-GA, Sequential Forward Selection, Simulated Annealing, Sequential Backward Selection, Randomized Hill Climbing are used.
- Filter Method - Makes an independent evaluation based on data characteristics and depends on specific metrics to, before learning begins, rate and select the best subset. For that Correlation based Feature Selection (CFS) algorithm is normally used with Fast Correlation Based Feature Selection (FCFS), and Markov Blanket Filter method.
- Embedded Method - As part of the learning procedure, performs variable selection and it is usually specific to some learning machines. For this process decision tree, Naive Bayes, random forest, Support Vector Machine (SVM), and based methods are normally used in regularization techniques, etc.

C. CLASSIFICATION APPROACHES AND VALIDATION

In the literature, we find four kinds of approaches to Internet traffic classification: port-based approaches, payload-based approaches, ML-based and statistical approaches. We provide in Table 4 a comparison among these kinds of approaches, which we briefly describe in the following.

1) PORT-BASED APPROACHES

The oldest traffic classification method is the port-based approach. According to [11], this method uses the association of well-known TCP/UDP port numbers assigned by IANA with ports in the TCP/UDP header. It uses port numbers related to an application where the application is related to a specific port number [36], some examples are SSH traffic that relates to port 22, and SMTP to 25. Most applications use port numbers already ''known'' so other hosts can start communication.

During handshake, an identifier is placed in the communication channel, right in the middle of the network, awaiting for SYN packets. SYN packets have the destination port number and are used during the handshake on TCP. The application is recognized by the port number contained in the SYN packet. It becomes all possible because TCP is connection orientated. Traffic identification through port numbers is also used on UDP, even though this protocol does not have control packet in its connection.

Implementing this method is quite simple and quick, once it does not involve calculations and requires only the number of ports to identify the application. Although its easy implementation, this approach has limitations that have a huge negative impact on traffic classification. Protocols that use tunnels, random ports, and Network Address Port Translation (NAPT) cannot be identified by this approach [8]. One possibility to easily escape detection by this method is to use port 80, which is generally open for HTTP traffic.

Some other protocols that cannot be identified by port-based approaches are the telephony through Internet that uses encapsulated Session Initiation Protocol (SIP) on Real-Time Transport Protocol (RTP), which sometimes use random port numbers, and P2P protocols that use random ports or ports associated to other protocols aiming to mask the traffic [8].

2) PAYLOAD-BASED APPROACHES

This kind of approaches recognizes applications by analyzing payload or packets. Aiming to find pre-defined byte sequences from the applications, payload is analyzed bit by bit. After those sequences, called signatures, are found, they are stored and compared to application packets for classification [36]. The great advantages of these methods are their capacity to generate low rates of false negatives and a highly accurate traffic classification.

TABLE 3. Summary of packet-based features.

Type	Features
Packet-Length Based	Packet length in bytes; packet size; packets number; Statistical values of packet size; Sequences of cumulative length; Relative frequency; Statistical features (mean, maximum, minimum, variance, standard deviation, median absolute deviation, percentiles, kurtosis, skew).
Packet-Ordering Based	Count the number of bursts; In each flow, mean and maximum burst length.
Packet-Timing Based	Packet inter-arrival time; Inter-packet delays.

TABLE 4. Comparison of the main approaches for traffic classification.

Kind of Approaches	Short Explanation	Limitations
Port-based	Associates port numbers to match applications.	Does not solve random or unknown port numbers.
Payload-based	Searches for protocol/application signatures in the form of string(s) in packet payload.	Cannot scan encrypted packets, encrypted payload and encrypted connections.
Statistics-based	Uses statistical values from the network or transport layers.	Generally does not specify application/client type.
ML-based	Automated method that foresees and makes a decision based on data analysis.	Set of pre-classified (also called pre-labeled).

The biggest limitations of these methods are: The development and maintenance of a database with application signatures. The high consumption of computational resources for the development requiring a longer processing time and storage space. It is an inefficient method to identify and classify traffic and packet payloads that are encrypted, unavailable payloads or on recognizing applications that have not been mapped yet. Besides that, it involves legitimacy and privacy issues of packets and traffic [54], [55].

Approaches based on statistical characteristics for traffic classification have been developed aiming to overcome limitations presented by traditional approaches, and they have caught the attention of researches. To identify and classify traffic, neural network and machine learning algorithms have been used.

3) ML-BASED APPROACHES

Machine Learning is known by supplying computers with the capacity to learn through programming. It has been used to prepare machines to work with data in a more efficient way. Machine learning is divided into unsupervised and supervised. On the unsupervised learning, information is extracted through non labeled data. On the other hand, on supervised learning, the information depends necessarily on data lettering. Machine Learning uses data patterns to label things [35], [42], [56].

ML has the capacity to work and learn from big data volume by using specific algorithms. Tasks as prevision, regression and classification of massive quantities of data can be solved through it. Machine Learning also has the capacity to deal with long and wide data. Long data means that number of subjects exceeded the number of input variables. Wide data corresponds to the number of input variables exceeding the numbers of subjects [35], [57], [58].

As appointed by [56], ML has a different and specific algorithm to solve problems involving data. Choosing the best algorithm to be used depends on which modal will better suit the problem, what the problem is and the quantity of variables involved in it.

Besides Internet traffic classification, ML has also been used in network operations and management, aiming to optimize the resources and improve the system performance. In addition, ML can be applied to many different areas such as marketing, games, digital images, intruders and malware detection, information security and data privacy.

4) STATISTICAL APPROACHES

Statistical-based classification uses statistics from the network and transport layers. By using parameters undependable of payload and payload analysis, statistical based classification methods go around payload, encrypted payloads and user privacy problems. They use statistical properties unique of protocols, flows and applications, which helps to differentiate the applications [36], [40].

Some examples of valid parameters of statistical-based network classification: packet inter-arrival time, flow duration, packet size, among others [36], [40]. Besides those parameters, statistical characteristics of packet tracking are captured and used, such as Border Gateway Protocol (BGP) updates and the unexpected rise of packet rate, which can also be an indicative of P2P applications in the network.

Commonly, Machine Learning uses statistical-based strategies to calculate resource parameters that will be used as data input in the supervised method classification, like SVM [36].

As stated by [59] techniques that are implemented based on statistical classification, are capable of perceiving flow behaviors expected through observations. Statistical methods

combined with methods grounded on rules might offer scalability, adaptability, flexibility and robustness. Furthermore, to differentiate traffic that has any flaws from regular traffic, statistical measurements can be used. However, the manual selection of statistical resources can compromise the requirements of traffic classification, generating a lower accuracy.

D. VALIDATION

The validating process consists on testing the obtained results from the classification, aiming to acquired its performance. In this sense, obtained classification results are compared to previously-known hand-based real data classification results, usually known as ground truth, which allows to compute true positive, false positive, true negative and false negative rates. Another challenge during validation is to collect original data in real time to obtain the ground truth [60].

Many performance measures are used to evaluate if a classification method could achieve the expected performance. Table 5 represents an overview of the metrics used to evaluate traffic classifiers. Metrics widely used are: F-measure [61], Precision, Recall, Specificity, Area Under Curve (AUC), Completeness [52], and F-1 Score [62].

III. STATISTICAL METHODS

In this section, we address the concept and properties of the statistical distances and divergence, as well as the SVM method based on statistics and widely used in traffic classification. Table 8 presents an overview of distances and divergences for quantitative (non-negative) data. We group the methods according to parametric and non-parametric approaches.

A. OVERVIEW OF PARAMETRIC AND NON-PARAMETRIC MODELS

On parametric models, datasets can be constructed by a probability distribution that has a number or a fixed set of parameters, which only the applied to variables. It is considered to be a parametric model some statistical and learning models that use a quantity of fixed parameters. For parametric ML, the quantity of parameters if fixed does not matter the amount of training data. Some examples of parametric models are linear SVM, Pearson correlation, denominated correlation information and Euclidean Distance [18], [63], [64], [65].

Non-parametric modals represent data without a defined number of parameters, and when modeling this data, they do not make presumptions about the probability distribution. Models implemented with this approach do not accept a specific mapping function between input and output data as true. This kind of models assumes that parameters are not only adjustable, but can also be altered. Parametric model also assumes that the larger the quantity of training data is, the larger will be the number of parameters. The result of this is that the non parametric model can take longer to perform the training [18], [65], [66]. Table 6 presents a comparison between parametric and non-parametric models. Bhattacharyya Distance, Hellinger Distance, KL Divergence, Wootters Distance, KS and Chi-square tests, and non-linear SVM are examples of non-parametric models. Hereafter the following subtopics will be approached: 1) Statistical Distances and 2) Statistical Divergences.

This section focus on Statistical Distances and Divergences. A brief description about these kinds of methods follows. Details about other methods herein mentioned that do not fall within those kind of statistical methods may be found elsewhere, namely details about Correlation Information (Pearson correlation) can be found in [67] and [68], details about Kolmogorov-Smirnov and Chi-Square tests can be found in [67], [69], and details about Shannon entropy can be found in [70] and [71].

1) STATISTICAL DISTANCES

The concept of distance between objects or individuals allows us to interpret, geometrically-wise, many classical techniques of multivariate analysis, equivalent to representing these objects as points in a metric space. In classification [72] of network traffic, the main objective is to find statistical differences between flows or even a pattern in traffic characteristics through statistical properties. It is possible to interpret this way because the observed variables are considered of a more general category, and not only as quantitative variables or own variables. As it is, it makes sense to calculate the proximity between objects or individuals [72], [73].

As stated by [72] the distance calculation is vital to many statistical inferences being them theoretical or applied. Besides that, it has become essential to solve data processing problems, such as classification, estimation, detection, regression, selection models, diagnosis, identification, recognition, indexation and compression. Combining its properties to statistical distance concepts, we have an essential instrument for science and data analysis [74].

Through the distance computation, it is possible to create hypotheses tests, study the estimators properties, compare classes, objects and individuals. Furthermore, the distance offers the researcher an assistance to interpret the data, because it is a very intuitive concept, allowing an easy comprehension and a harmonious representation [74], [75].

In general, we consider two classes of statistical distances between individuals and populations. The individuals of each population are characterized by a random vector $X = (X_1, \ldots, X_P)$, which follows a probability distribution $f(x_i, \ldots, x_p; \theta)$. The distance between two individuals i, j, characterized by the points x_i, x_j, of R^p, is a non-negative symmetric measure, $\delta(x_i, x_j)$, which will depend on θ, where θ represents the parameters and R^p is the quantity of dimensions that the X variable may have. Therefore X has n observations and p variables.

Moreover, the distance between two populations will be measured by the divergence $\delta(\theta_1, \theta_2)$ between the parameters that characterize them. It may also be convenient to enter the distance $\delta(x_i, \theta)$ between an individual i and the θ parameters. Non-parametric distances can be defines by it functional

TABLE 5. Summary of the metrics often used to evaluate the traffic classifiers, where TP means True Positive, TN means True Negative, FP means False Positive, FN means False Negative, TPR means True Positive Rate, TNR means True Negative Rate.

Metrics	Description	Definition
Accuracy (A) [61]	It is the ratio between cases classified as truly positive and negative and the sum of all positive and negative cases predicted in the classification.	$\frac{TN+TP}{FP+TN+FN+TP}$
Precision (P) [61]	It correctly evaluates how many cases are identified as positive.	$\frac{TP}{FP+TP}$
Recall (R) [61]	It is known as true positive rate or hit rate, it presents the rate of positive cases that were correctly identified by the classifier in the dataset.	$\frac{TP}{FN+TP} (= TPR)$
Sensitivity [52]	It is also known as Recall metric.	$\frac{TP}{FN+TP}$
Specificity [62]	It calculates the number of correctly classified positive cases for the total positive cases found.	$TNR = \frac{TN}{TN+FP}$ or $\frac{TN}{TN+FP} = 1 - FPR$
Completeness [52]	For the total number of positive cases, the proportion of correctly or incorrectly classified positive cases is measured.	$\frac{FP+TP}{FN+TP}$
F-Measure [61]	It measures the effectiveness of debug testing, it is considered harmonic calculation between Precision and Recall.	$\frac{2*\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$
F1-Score [62]	It is the harmonic mean between Precision and Sensitivity.	$\frac{2TP}{2TP+FP+FN}$
Area Under the Curve (AUC) [52]	It is known as Receiver Operating Characteristics (ROC)	$\frac{1+TPR-FPR}{2}$ or $\frac{\text{Sensitivity} + \text{Specificity}}{2}$
False Positive Rate (FPR) [62]	It is the calculation of the rate of negatives incorrectly classified as positives	$\frac{FP}{N} = \frac{FP}{FP+TN}$
False Negative Rate (FNR) [62]	It is the calculation of the rate of positives incorrectly classified as negatives	$\frac{FN}{P} = \frac{FN}{FN+TP}$
Geometric mean (G-mean) [62]	It is the calculation of the correlation between the rate of positives and the classified results	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$

TABLE 6. Side-by-side comparison of parametric and non-parametric classifiers.

Parametric classifier	Non-Parametric classifier
The model is built using a fixed number of parameters.	The model is built using a flexible number of parameters.
It can only be applied to variables.	It can be applied to Attributes and Variables.
It makes strong data assumptions.	It normally does not make data assumptions.
It needs less data.	It needs much more data
It assumes a normal distribution.	No distribution is assumed.
Data manipulation – Ratio or interval data.	Original data is manipulated.
Outliers can seriously effect the results.	Outliers cannot seriously affect the results.
Performance peaks when the spread of each group is different.	Performance peaks when the spread of each group is the same.
It has more statistical power.	It has less statistical power.
It is faster, computationally speaking.	It is not so fast when compared to parametric models

divergence and the density functions. In some cases they are related to entropy measurements.

A δ distance over an Ω set is an application of $\Omega \times \Omega$ over R so that each pair (i, j) corresponds to a real number $\delta(i, j) = \delta ij$ fulfilling some of the following properties, according to the Table 7.

A distance must fulfill at least properties 1, 2, 3, presented in Table 7. When it fills these properties, it is called dissimilarity. In general, δ only meets approximately some of the stated properties. It is then a matter of representing (Ω, δ) through a model (V, d), approximating δ to d, where δ meets sufficient properties that are mandatory.

TABLE 7. General properties of distances and divergences and qualification of a distance according to its properties, where δ_{ij} represents the distance between pairs.

Qualification of a distance according to its properties	Distance property
Dissimilarity: 1, 2, 3	$1 - \delta_{ij} \geq 0$
Metric distance: 1, 2, 3, 4, 5	$2 - \delta_{ij} = 0$
Ultrametric distance: 1, 2, 3, 6	$3 - \delta_{ij} = \delta_{ij}$
Euclidean distance: 1, 2, 3, 4, 8	$4 - \delta_{ij} \leq \delta_{ik} + \delta_{ik}$
Additive distance: 1, 2, 3, 7	$5 - \delta_{ij} = 0 \Leftrightarrow i = j$
Divergence: 1, 2, 10	$6 - \delta_{ij} \leq max\delta_{ik}, \delta_{ik}$ (ultrametric inequality)
	$7 - \delta_{ij} + \delta_{kl} \leq max\delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk}$ (additive inequality)
	8 - δ_{ij} is euclidean
	9 - δ_{ij} is riemannian
	10 –δ_{ij} is a divergence

According to the representation technique, such as main component analysis, main coordinate analysis, proximity, correspondence analysis, cluster analysis, the distance d can be Euclidean, ultrametric, additive, non-Euclidean, or Riemanian, among others.

2) STATISTICAL DIVERGENCES

Non-parametric measures of divergence between probability distributions are defined as functional expressions often related to information theory, which measures the degree of discrepancy between any two distributions, not necessarily belonging to the same parametric family. Divergences have applications in statistical inference and in stochastic processes.

Let $p = (p_i, \ldots, p_n)$, $q = (q_i, \ldots, q_n)$ be two multinomial distributions. The divergence between q and p can be measured as the discrepancy between the quotient $x_i = q_i/p_i$ and 1. Based on the meaning of $Ho = (p) - entropy$, $\phi-$ Csiszar divergence is defined between p and q, where ϕ is a strictly convex function in which $\phi(1) = 0.H\phi$, and by Jensen inequality we have:

$$C_\phi[p, q] = \sum p_i \phi(x_i) \geq \phi(\sum p_i x_i) = \phi(1) = 0. \quad (1)$$

The equation 1 reaches the value 0 if and only if $p = q$. It can be taken as a measure of dissimilarity between p and q, but in general it is not a distance, as it is not always symmetrical, or if it is, it may not meet the triangular inequality. Shannon entropy and the $\phi-$ Csiszar divergence form the information measure known as the KL [25].

B. PARAMETRIC DISTANCES AND DIVERGENCES

1) EUCLIDEAN DISTANCE

The most familiar distance between two individuals i, j is the Euclidean distance described by the equation [25]:

$$D_E[i, j] = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - y_{jk}\right)^2}. \quad (2)$$

Proposed by the Greek mathematician Euclid, it is based on calculating the distance between two points within the Euclidean space. where $D_E[i, j]$ represents the distance function, p defines the quantity of samples, k defines the initial value of the sample, x_{ik} represents the first point and y_{jk} represents the second point [76].

2) JENSEN-SHANNON DIVERGENCE

Jensen-Shannon Divergence (JSD) is the calculation of the difference between two series of probability distributions [77]. It is known for being the limited symmetrization of KL [78].

JSD is a function that allows us to quantify the difference of two, maybe more, probability distributions [22]. JSD also has the additional advantage of not requiring absolute continuity of the distributions to compare them. Thereby, JSD can be used to compare the distribution of different packet sequences in a network flow, associating an appearing frequency to each flow with probability distribution.

For two discrete probability distributions $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$ with $pi \geq 0$, $qi \geq 0$, JSD divergence is represented by [77]:

$$JSD\,(P, Q) = \frac{1}{2}\left\{\sum_{i=1}^{N} pi \log\left(\frac{2pi}{pi + qi}\right) + \sum_{i=1}^{N} qi \log\left(\frac{2qi}{pi + qi}\right)\right\}. \quad (3)$$

JSD function equals 0, if and only if $(pi = qi)$. In this case, it means that they are the same distribution, in other words, the same application. It is a delimited and symmetric metric $(0 \leq JSD \leq log(2))$ for orthogonal distributions $(pi.qi = 0)$. As traffic classification was intended through the values of the distances between the application distributions, JSD determines the divergence between two probability distributions P and Q.

C. NON-PARAMETRIC DISTANCES AND DIVERGENCES

1) BHATTACHARYYA DISTANCE

Bhattacharyya Distance, also known as divergence, was proposed by a statistician called Anil Kumar Bhattacharyya

(1943 and 1946) working with Kailath [79]. This distance measures the dissimilarity between two probability distributions. It is very related to Bhattacharyya coefficient, that is the calculation of the quantity of overlap of two statistical population samples [80], [81]. In its first version, Bhattacharyya did not present the calculation, he used a logarithm scale.

Bhattacharyya Distance is independent of the distribution function and it can be applied to any data set or sample. This characteristic makes the distance appealing to be used in models in which the distribution is undetermined [80].

Bhattacharyya coefficient can be used in classification as a measure of the separability between classes [82], and to determine the relative proximity between samples that are being taken under consideration.

When two probability distributions have similar averages, Bhattacharyya Distance rises depending on the difference between standard deviations, in other words, the bigger the difference between standard deviations, the bigger the probability distribution. Bhattacharyya statistical distribution is given by equation 4 [83]:

$$Bc_D(P, Q) = -\log\left(\sum_{i=1}^{N}\sqrt{pi \times qi}\right), \quad (4)$$

where N is the quantity of partitions and pi and qi is the quantity of members from the sample in the I-th partition.

2) HELLINGER DISTANCE

Hellinger Distance was proposed by the German mathematician Ernst David Hellinger in 1909. It is a statistical divergence used to calculate the dissimilarity between two probability distributions. Hellinger Distance (HD) is related to Bhattacharyya Distance and it is part of the f-divergences family [78].

Studies presented in [84] and [85] showed that Hellinger Distance can be used in classification. On the current scenario, this distance has been very used in machine learning, even as an alternative to methods such as entropy, aiming to detect failures in the classifiers [86] and breakpoints on the performance of those classifiers [87]. Furthermore, according to the literature, Hellinger Distance has been used in many parametric models being very successful on solving problems of statistical estimation [84], [85]. The calculation function is obtained from two probability distributions p and q as follows [85]:

$$H_D(P, Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{N}\left(\sqrt{pi} - \sqrt{qi}\right)^2}. \quad (5)$$

Hellinger Distance is non-negative and symmetric, and $H_D(P, Q)$ is in $\left[0, \sqrt{2}\right]$. Note that the higher Hellinger Distance is, the better the differentiation between probabilities will be.

3) KULLBACK-LEIBLER DIVERGENCE

KL Divergence, well known as relative entropy, was defined by the mathematicians Solomon Kullback and Richard A. Leibler in 1951. It represents the calculation between two probability distributions [88], [89], [90], [91]. Through statistical testing, those mathematicians started from the principle that two probability distributions are different, since there is a possibility of differentiation between them. KL measures the information gain and has been used in statistics, specially in Bayesian statistics.

KL is considered a special class of divergence, being an asymmetric measurement of difference or not dissimilarity. Therefore KL allows us to deduce both the difference and the not dissimilarity between two distributions [91]. In KL, p_i e q_i are considered probability distributions, where the function is represented by $D_{kl}[p||q]$.

$$D_{kl}[p||q] = \sum pi \log\left(\frac{1}{pi}\right) - \sum pi \log\left(\frac{1}{qi}\right). \quad (6)$$

It can also be given by the equation:

$$D_{kl}[p||q] = \sum pi \log\left(\frac{pi}{qi}\right). \quad (7)$$

On problems of data processing or classification, the result of the function $D_{kl}[p||q]$ is the calculation of the expected p value, essential on samples based on q. Normally, the data is represented by p that assumes the real or current distribution of class, flow, application or model that are represented by the q variable [92], [93].

$$D_{kl}[p||q] = \sum_{i=1}^{N} p(x_i) log \frac{p(x_i)}{q(x_i)}, \quad (8)$$

where N defines the quantity of samples. See that the symmetric version of KL Divergence is the Jensen-Shannon Divergence [78], [94].

4) WOOTTERS DISTANCE

Wootters Distance was proposed by the American physicist William Wootters in 1981, aiming to calculate the probability differences under the values of typical fluctuations. The main idea of this distance is to properly consider the statistical fluctuations inherent to any finite sample. It is purely and simply statistical and the concept can be used in any probabilistic area [95].

Considering two probability distributions p and q, the minimal distance between two points will be equivalent to the angle presented by them, represented by the equation 9 [83], [95]. Wootters can also define the not dissimilarity between two samples [96]. Given two probability distributions $P_i = \{p_j^{(i)}, j = 1, \ldots, N\}$ with $i = 1, 2$.

$$Wo_D(P, Q) = \arccos\left(\sum_{i=1}^{N}\sqrt{pi \times qi}\right). \quad (9)$$

Note that *arccos*() decreases in [0, 1], and that distances were used to discriminate traffic. Table 8 presents a summary of distances and divergences for quantitative (non-negative) data.

TABLE 8. Summary of distances and divergences for quantitative (non-negative) data.

Distance / Divergence	Explanation
Bhattacharyya Distance [83]	$Bc_D(P,Q) = -\log\left(\sum_{i=1}^{N}\sqrt{pi \times qi}\right)$
Euclidean Distance [25]	$D_E[i,j] = \sqrt{\sum_{k=1}^{p}(x_{ik} - y_{jk})^2}$
Hellinger Distance [83]	$H_D(P,Q) = \sqrt{\frac{1}{2}\sum_{i=1}^{N}(\sqrt{pi} - \sqrt{qi})^2}$
Jensen-Shannon Divergence [77]	$JSD(P,Q) = \frac{1}{2}\left\{\sum_{i=1}^{N} pi \log\left(\frac{2pi}{pi+qi}\right) + \sum_{i=1}^{N} qi \log\left(\frac{2qi}{pi+qi}\right)\right\}$
Kullback–Leibler Divergence [88], [89], [90], [91]	$D_{kL}[p\|\|q] = \sum_{i=1}^{N} pi \log\left(\frac{qi}{pi}\right)$
Wootters Distance [95]	$(P,Q) = \arccos\left(\sum_{i=1}^{N}\sqrt{pi \times qi}\right)$

D. SUPPORT VECTOR MACHINES

SVM was developed by Vapnik, Guyon, and Hastie [97], based on the Statistical Learning Theory and aims to solve pattern classification problems. Statistical Learning Theory gives us mathematical conditions to choose an efficient classifier to train and test a specific set of data. SVM is a supervised method focused on classification and regression. To classify, initially, SVM was developed seeking binary classification capable of recognizing sample patterns in predefined classes [98].

Currently, SVM supports the task of multi-class learning and it is used to solve problems such as multi-classification. In addition, it has been widely used in the field of artificial intelligence. SVM is responsible for finding the best possible separation boundary between classes/labels for a given set of data that is linearly separable. For SVM, the many separation boundaries that are capable of completely separating classes are called hyperplanes. A decision plane that separates a set of objects with different class members is a hyperplane [99].

An important SVM aspect is the margin, which is seen as a breach between the two lines closest to the class points. The margin is calculated as the perpendicular distance of the support points closest to the vectors. A good margin is the one which has the greatest distance between classes, a lowers margin is a bad one [99].

SVM seeks to find the best hyperplane for a given data set whose classes are linearly separable. SVM builds a classifier according to a set of patterns identified by it in the training examples [100].

Classification problems tend to be more elaborated, requiring optimal separation through more complex structures. SVM proposes the classification of new objects (test) based on available data (training). For that, a set of mathematical functions is used to map the new objects, known as Kernels. SVM kernels are divided into two versions, linear and non-linear [92].

Kernel functions are intended to project vectors of input feature into a high-dimensional feature space to classify issues which lie in non-linearly separable spaces. This is done because as the problem of dimensional space increases, the probability of this problem becoming linearly separable around a low-dimensional space also increases. However, to obtain a good distribution of the complex problem, a training set with a high number of instances is necessary. SVM-based classification uses kernel functions Linear, Radial Base Function kernel(RBF), Polynomial, and Sigmoid [29], [101].

- Linear: it is the scalar product of observations. It is the sum of the multiplication of every pair of input vectors.
- RBF: it maps an input space in a finite dimensional space. It is the most used Kernel in SVM classification.
- Polynomial: This kernel distinguishes a non-linear input space from a curved one. It is known for being more generalized than linear kernel.
- Sigmoid: Neural networks use the sigmoid kernel as the activation function. This kernel is part of the class of differentiable, limited and crescent monotonically functions.

Note that SVM-based classification kernel function Linear is considered a parametric model, while the kernel functions RBF, Polynomial, and Sigmoid are considered a non-parametric models.

RBF and Polynomial are both suggestive kernels to separate non-linear application classes from curved ones. Through this choice, more precise classifiers can be obtained. RBF and Polynomial Kernels calculate the separation line in the higher dimension to classify some applications.

An important thing about SVM is the regulation parameters that can be used to configure the SVM [29]. One of them is the C parameter, which is the penalty parameter that represents the classification error or the error term, and it is used to maintain the regulation of the model. SVM optimization depends on controlling how much error can be handed. It is this way that trade-off is controlled between incorrect classification terms and the decision limit. See that the lower the value of C, the lower hyperplane margin and the greater the value of C, the greater the margin will be [102].

Another parameter that also deserves attention on SVM is the Gamma parameter. Low values of Gamma parameter makes so the data does not adapt much to the training data set. Now when the values are higher, the data adapts perfectly to the training set. See that there must have a balance on Gamma values, because values too high can cause an over adjustment and values too low may consider only points close to the margin.

IV. CLASSIFICATION OF INTERNET TRAFFIC USING STATISTICAL METHODS

This section addresses the use of statistical methods for Internet traffic classification. Tables 9 to 12 present an overview of previously used statistical methods for Internet traffic classification, as well as their main characteristics and performance. In these tables, the performance values are given in %. The metrics are indicated as follows: Recall (R), Accuracy (A), Precision (P), F-Measure (FM), F-score (F1), Area Under the Curves (AUC), Receiver Operating Characteristic (ROC), False Positive Rate (FPR), True Positive Rate (TPR), Kappa (K), Geometric mean (G-mean), Specificity (Sp), Sensitivity (S), Not Available (N/A).

To the best of our knowledge, Wootters Distance, addressed in the previous section, has not yet been investigated for Internet traffic classification. Therefore it is not considered in this section, being left as a possible future research direction in the next section. Some methods have few applications and were little explored for classification, such as Jansen-Shannon and KL. Others were quite explored, such as SVM, which has been extensively explored in this kind of classification, often presenting good accuracy values.

A. DISTANCE-BASED METHODS

Table 9 details the papers describing distance-based methods for traffic statistical analysis.

1) EUCLIDEAN DISTANCE

Euclidean Distance was addressed in several works found in the literature, including on the implementation of some famous machine learning algorithms, such as K-mean, and Nearest Neighbor (NN). In Table 9, there is a summary of works around this distance. Zhu et al. in [106] proposed a method for classifying an unknown protocol of the application layer based on the Euclidean Distance. In [55], Shi et al. discuss the method of extraction and selection of features for classification, where K-Means algorithm with Euclidean Distance were used to group the features. Pereira et al. in [103] developed a network traffic classification system based on real-time flow using NN technique and Euclidean Distance. The focus of [105] was to use statistical resources of network flows to identify the generated application, and the Euclidean Distance was used to test the classification algorithm. Singh in [104] used K-Means which calculates the distance between objects by using the Euclidean distance to group the network traffic applications.

2) BHATTACHARYYA DISTANCE

Shah and Dang in [114] used Bhattacharyya Distance to select the the highest distance features from a test pool. In [110], the temporal analysis of the behavior of the network is established by calculating this same distance. Aiming to calculate the difference among solved and unsolved iEvents that correspond to the traffic density distributions, Zanin [107] also used this distance. In [111], Class separability was maximized using the Bhattacharyya Distance algorithm. In [108], the Bhattacharyya Distance is used to quantify the not dissimilarity of the probability distributions of Virtual Machine (VM) resources usage. In [109], the Bhattacharyya Distance is used to calculate changes of color histogram. In [112], Baskoro et al. proposed an algorithm for counting and tracking vehicles using the Bhattacharyya Distance. It is used by Laz in [113] to evaluate detection system performance. In Table 9, there is a summary of works around Bhattacharyya Distance.

3) HELLINGER DISTANCE

In Table 9, there is a summary of works about the use of Hellinger Distance. In [118] the Hellinger Distance was used by Wang et al. to find the deviations among sketches. A sketch is a collection of hash tables where Wang et al. propose the SkyShield method using the sketch technique aiming to detect anomalies. The Hellinger Distance was used in [121] to perform linear and non-linear transformations aiming the improvement of accuracy in dataset classification. Derivation of the Hellinger square distance was used by Liu et al. in [115]. In [119] Kumari and Thakar proposed an oversampling method based on the Hellinger Distance to identify the minority class in the classification. In [116] it is used to measure the not dissimilarity of two probability distributions to implement an attack classifier in a monitoring network. It was also used in [117] on the linear SVM kernel implementation for the classifier training step. In [120] Hellinger Distance is used on feature value distribution.

4) WOOTTERS DISTANCE

In the research made throughout the databases referred to in this article during the period from 2011 to 2022, applications of Wootters Distance as a classification technique, feature selection and kernel increment in methods such as SVM, for example, were not found in the literature.

B. DIVERGENCE-BASED METHODS

Table 10 details the papers describing divergence-based methods for traffic statistical classification.

1) JENSEN-SHANNON DIVERGENCE

In Table 10, there is a summary of works around the use of JSD. In [123] Zareapoor et al. applied JSD property to identify information deviation. In [124], Zhi et al. proposed an Interest Flooding Attack (IFA), that consists of a resistance mechanism based on JSD. This mechanism can help detect and mitigate Flooding Attack on the network. The obtained values from the JSD calculation were used on [125] to select the features. In [126] JSD was used to calculate the distribution not dissimilarity among original discrete attributes and the generated ones, aiming to evaluate the Anti-Intrusion Detection Autoencoder (AIDAE) performance. In [122], the difference between M1 and M2 (the histograms of two mixture distributions) is quantified using JSD of bin-placement approaches.

TABLE 9. Works related to distance-based statistical methods (Euclidean Distance, Bhattacharyya Distance, Hellinger Distance).

Method	Work	Year	Characteristics	Performance
Euclidean Distance	Pereira *et al.* [103]	2015	Features: number of packets, number of bytes, elapsed time between the first and last packets, the number of all packets with at least a byte of TCP data payload, the median and the variance of the number of bytes in IP packet, and the number of all packets seen with the PUSH bit set in the TCP header. Applications: HTTP and HTTPS, FTP, WWW, XVTTP, and ISAKMP.	A:87.40-89.86
	Singh [104]	2015	Features: packet length and inter-arrival time including (average, maximum, minimum, and standard deviation), number of bytes transferred, total number of packet in flow and Flow duration. Applications: HTTP, DHCP, ICMP, DNS, and SMTP. Technique: correlation-based feature selection-CFS.	A:55.00-88.00
	Shi *et al.* [55]	2017	Features: extract the multifractal features, multifractal spectrum, largest wavelet coefficient, variance ratio, cumulative variance ratio. Applications: P2P, WWW, flash+HTTP, IM, SMTP, VoIP, IMAP and POP. Techniques: method of linear regressions, and Wavelet Leaders Multifractal Formalism (WLMF).	A:55.70-99.80
	Schmidt *et al.* [105]	2017	Features: number of pushed data packets, median of total bytes in IP packets, port number at server, bytes in the initial window, average segment size, bytes in the initial window, packets with at least a byte of TCP data payload, the total number of Round Trip Time (RTT) samples, variance of bytes in Ethernet packet, packets with the PUSH bit set in the TCP header, and the minimum segment size. Aplications: Postgres, FTP, Oracle, Sqlnet, IMAP, SSH, SMTP, POP2/3, X11, WWW, LDAP, DNS, KaZaA, NTP, BitTorrent, Games, Windows Media Player, and Worm and virus attacks. Techniques: Manhattan Distance, Euclidean, Chebyshev distance, and Cosine Distance.	A:88.00-94.77, FM:86.90
	Zhu *et al.* [106]	2019	Features: number of labeled protocol in the dataset, protocol flow statistics, longest distance, average distance. Applications: SMTP, HTTP, FTP, Bittorrent, and POP3. Techniques: clustering, and deep neural network.	A:96.00
Bhattacharyya Distance	Zanin [107]	2013	Features: analysis of the statistical properties, and Data Science analysis.	N/A
	Canali and Lancellotti [108]	2013	Features: statistical properties of Virtual Machine.	N/A
	Dinani *et al.* [109]	2015	Features: overall mean, averaged in a given time duration of video, standard deviation, skew, R-inverse variance, uniformity, pixel length, and entropy.	N/A
	Sadreazami *et al.* [110]	2017	Features: signal statistics, mean, variance, and time.	ROC:93.15-99.75
	Sameen and Pradhan [111]	2017	Features: spectral, spatial, and texture properties. Technique: fuzzy logic for define rules.	N/A
	Baskoro *et al.* [112]	2017	Features: Probability Density Function (PDF), number of pixel, and color pdf's.	P:96.5, R:96.3
	Laz [113]	2017	Techniques: lagrange multipliers technique, and parallel computing.	N/A
	Shah and Dang [114]	2019/ 2020	Features: probability distribution, modulation pairs, and maximum distance.	N/A
Hellinger Distance	Liu *et al.* [115]	2014	Features: exponential distribution, Erlang distribution, small average distribution distance, and maximum entropy.	N/A
	Safarik *et al.* [116]	2014	Applications: SIP message, SIP attack classification, IP addresses, and specific SIP header values or ports.	N/A
	Luo *et al.* [117]	2015	Features: number of foreground pixels, number of background road pixels, and density ratio. Techniques: regression models, Pearson, and correlation coefficient.	A:83
	Wang *et al.* [118]	2017	Features: number of hash functions, size of hash tables, and probability vector.	FPR:0-35, TPR:38-80
	Kumari and Thakar [119]	2017	Features: probability distribution, and synthetic sample value.	AUC: 69-94
	Liu *et al.* [120]	2019	Features: Packet size sequences, and inter-arrival time. Applications: Social, Streaming, Web, and Download.	A: 77.77, G-mean:0-90

2) KULLBACK-LEIBLER DIVERGENCE

Some works were found in the literature using KL for Internet traffic classification. In Table 10, there is a summary of works about the use of this divergence. Kim et al. in [127] proposed a network classification with a KL criterion. In [128], it was used to detect video clips. KL was also used in other fields of analysis, such as agriculture. In [129] KL is employed to validate the not dissimilarity of unknown pixels. In [130], KL is used to classification of encrypted internet traffic.

C. SVM

Several SVM applications for traffic classification were found in the literature. In Table 11, there is a summary of works around this statistical method. It was used in [133] with the linear, Polynomial, Sigmoid and Radial kernels for traffic classification on a Software Defined Networking-SDN. Cao et al. in [134] proposed a real-time training model using SVM. It was also used in [139] with denoising schemes to improve prediction accuracy. In [136] Miao et al. used SVM to optimize feature selection. To distinguish data representing normal network traffic and Distributed Denial of Service (DDoS) flows, Aamir and Zaidi [140] tested different combinations of parameters on SVM. In [143], Sentas et al. developed a video data detection and classification system. Luo et al. in [141] proposed the Least Square SVM (LSSVM) hybrid optimized, a model for short-term traffic flow forecasting. Suresh and Srijanee in [145] used

TABLE 10. Works related to divergence-based statistical methods (Jensen-Shannon Divergence, Kullback–Leibler Divergence).

Method	Work	Year	Characteristics	Performance
Jensen-Shannon Divergence	Garcia and Korhonen [122]	2018	Features: total amount of Bytes in packets, number of packets in a flow, time between first and last packet sizes, min/max of packet sizes, number of downlink packets, skew/kurtosis of packet sizes, mean of packet sizes, and standard deviation/variance of packets. Technique: Random forest.	A:96, P:88, R:95, ROC:94
	Zareapoor *et al.* [123]	2018	Technique: JSD	N/A
	Zhi *et al.* [124]	2019	Features: probability distribution, high entropy values signify a more dispersed probability distribution, and number of data packets sequential time interval.	ROC:98.88
	Barut *et al.* [125]	2020	Features: distribution of each feature, average, length numerical, variable sizes, the mean value of array, the length of the array, the maximum value in array, the minimum value in array. Techniques: correlation, random forest algorithm, and Principal Component Analysis (PCA).	R:51-93, P:0-100, FM:69-90
	Chen *et al.* [126]	2020	Features: Mean Square Error (MSE), distribution of continuous features, and number of discrete features.	N/A
Kullback-Leibler Divergence	Kim *et al.* [127]	2016	Feature: maximum sequence size-MSS value. Applications: IMAP, and SMTP. Techniques: markov model, and concept of bag based on port number.	R:20-70, P:39.13-99.85
	Xu *et al.* [128]	2016	Features: stationary stochastic, probability distribution, discrete optical flow approach, number of frames in a video, and length sequence. Application: video. Techniques: bag-of-words paradigm, MPEG motion vectors, and Fourier coefficients.	AUC:59.43-78.80
	Zhang *et al.* [129]	2019	Features: reference time series data, and probability distribution of the NDVI. Application: video.	A:94.80, K:85
	Cunha *et al.* [130]	2020	Feature: Relative frequency. Applications: HTTP and Flash-based, RTSP, MMS, P2P streaming, PPStream, TVUPlayer and SopCast, P2P file-sharing: BitTorrent, e-Donkey, and Gnutella, VoIP: Skype, Google Talk, SIP traffic, FTP and SFTP transfers, Telnet and SSH sessions. Techniques: KL Divergence calculation, and heuristics.	A:99-100, P:100, FM:85-100, R:74-100

SVM to analyze the traffic data pattern and detect anomalies in order to secure high-volume confidential data transmitted over wireless network. In [142], Xiao used this statistical method combined with KNN to detect traffic incidents. In [144] Dong proposed optimizing SVM method to improve training speed and classification, using this enhanced SVM called Cost-Sensitive SVM (CMSVM) to solve imbalance in network traffic identification. Cao and Fang [146] and Syarif et al. [63] optimized the SVM parameters based on the Genetic Algorithm (GA) for Internet traffic classification. Mostafa et al. in [147] proposed a new version of this method named Relaxed Constraint Support Vector Machines (RSVMs) to optimize classification without needing source or destination IP addresses or port information. In [100] Liu et al. addressed SVM for Traffic Identification and Classification (STIC) aiming to identify applications, focusing on the duration and quality of YouTube streaming. Aggarwal and Singh in [135] made use of this method to categorize Internet traffic. In [148], a distributed SVM framework was implemented to classify network traffic using Hadoop. In [131], Hao et al. improve a variation of it called Directed Acyclic Graph-Support Vector Machine (DAGSVM) to classify network traffic. In [132], SVM was used to sort network traffic by improving the algorithm to calculate its own resource weights and parameter values for every individual binary classifier. It was also used in [138] to classify large amounts of data. SVM was used in [137] as the basis to implement an optimized model in order to reduce memory and CPU cost in the training phase, called Incremental SVM (ISVM), and a modified version with Attenuation factor (AISVM).

D. OTHER METHODS

Table 12 details the papers describing various other methods for traffic statistical analysis found in the literature.

1) CORRELATION INFORMATION

In Table 12, there is a summary of works around Correlation Information (Pearson Correlation). Correlation was used in [61] to boost network traffic ranking performance. In [135], Aggarwal and Singh used a Bag of Flow (BoF) to model correlation information in traffic flows and SVM to categorize traffic by application. The correlation was also object of research on [149], that presented a new traffic classification framework. For that, Zhang et al. used the BoF to model information of traffic flow correlation. Besides that they also used a model based on NN. A new classification method that took under consideration the network traffic flow correlation was also proposed by Zhang et al. in [149]. In [151], Zhang et al. considered real traffic and classified the correlated flows together. In Dong et al. [150] presented the disadvantages of using Pearson's Correlation Coefficient to measure the relationship between traffic flows. From the disadvantages, the authors presented a new proposal based on metric correlation quantitatively and accurately.

2) STATISTICAL KOLMOGOROV-SMIRNOV AND CHI-SQUARE TESTS

Statistics such as Kolmogorov-Smirnov and Chi-Square tests have also been used for traffic classification. In Table 12, there is a summary of works around those tests.

TABLE 11. Works related to SVM statistical methods.

Method	Work	Year	Characteristics	Performance
SVM	Hao *et al.* [131]	2015	Features: selection algorithm, and Chi-square values. Applications: Mail, WWW.	A:96.58
	Hao *et al.* [132]	2015	Features: total number of bytes sent by client to server, Fast Correlation-Based Filter (FCBF), feature selection algorithm - server port, maximum of bytes in Ethernet packet, average window advertisement, total number of bytes sent by server to client, average segment size, minimum window advertisement, minimum segment size, maximum segment size, and maximum of total bytes in IP packet. Applications: WWW, Mail, FTP-Control, FTP-PASV, Attack, P2P, Database, FTP-Data, and Multimedia, Services.	A:57.38 -97.00
	Syarif *et al.* [63]	2016	Features: Particle Swarm Optimization (PSO), and Feature selection algorithm- Genetic Algorithm-GA. Datasets: Embryonal Tumours, Leukemia, Dexter, Madelon, Internet_ads, Spambase, Musk, Intrusion NSL KDD, and SPECTF Heart.	FM: 76.67-95.68
	Fan and Liu [133]	2017	Features: mean segment size, round trip time, and packet inter-arrival time. Applications: Web, SMTP, POP3, IMAP, FTP, DNS, X11, NTP, BitTorrent, eDonkey, Mysql, Oracle, Windows Media Player, Virus, Worm, Telnet, SSH, and Games.	A:80.59-97.96, P:12.5-99.24, R:2.43-99.93, FM:4.08-99.58
	Cao *et al.* [134]	2017	Features: Feature dimension by principal component analysis (PCA), and Number of folds. Applications: Mail, WWW, Attack, FTP, Database, P2P, Services, and Multimedia. Technique: Correlation-Based Feature Selection (CFS).	A:11.81-99.90
	Aggarwal and Singh [135]	2017	Features: Probability Density Function (PDF), size of the first packets of an SSL, and statistical features. Application: P2P-TV traffic.	A:88.87
	Miao *et al.* [136]	2018	Features: bytes volume, packets quantity, packet size statistic information (Min.,Max., Ave. and variance), duration, and interpacket time statistic features. Applications: EBUDDY, DNS, eDonkey, HTTP, FTP, MSN, IMAP, SMTP, POP3, RSP, RTSP, SMB, XMPP, SSL2, SSL3, YAHOOMSG, and SSH. Techniques: NN, and RandomForest.	A:25.01-92.92, FM:6.14-99.70
	Liu *et al.* [100]	2018	Features: sequence of packets from a source, unidirectional flow, and bidirectional flow, and packets in a specific transport. Applications: Google page, Yahoo page, YouTube, Facebook, Line, BitTorrent, eDonkey, Skype, League of Legends, Twitter, Twitch, Messager, Google Hangout, Instagram, Spotify, Dropbox, OneDrive, KKBOX, MoPTT, Sanguosha, PPS, WooTalk, IRC, Garena Messager, Foxy, Pokémon Go, and QQ,	A:92.54-99.00, R:92.73-98.89, P:92.21-99.00, FM:92.23-98.89
	Sun *et al.* [137]	2018	Features: attributes of the traffic flow, dimension of features, packet size, packet length, inter-packet timing, TCP window size, and information derived from traffic flows. Applications: WWW, P2P, and FTP.	A:82.90-95.40
	Akinyelu and Absalom [138]	2019	Features: Wrapper-based technique, and Filter-based technique.	A:55.11-99.86
	Tang *et al.* [139]	2019	Features: sampling interval, distribution of denoised traffic flow, and Number of forecasting. Techniques: Empirical Mode Decomposition, Wavelet (WL), Ensemble Empirical Mode Decomposition (EEMD), ButterWorth (BW) filter, and Moving Average (MA).	N/A
	Aamir and Zaidi [140]	2019	Features: bwd packet length standard, cumulative entropies of clusters, flow duration, average packet size, and flow.	AUC: 95.04-96.75
	Luo *et al.* [141]	2019	Features: total sample size, true value at period, prediction value at period, particle swarm size, the maximum iteration number, cognitive factor, Social factor, and probability. Techniques: Root Mean Square Error (RMSE), the Equal Coefficient (EC), and Mean Absolute Error (MAE).	N/A
	Xiao [142]	2019	Technique: KNN	N/A
	Sentas *et al.* [143]	2020	Features: image size in pixels, block size, block stride, and block stride in pixel. Application: video, formats: by ImageNet. Technique: Region Of Interest (ROI).	R:78.81-98.55, P:87.73-98.55
	Dong [144]	2021	Features: high and low port number, flow transport protocol, and flow duration, TCP header flag including (TCPflags1, TCPflags2), bi-direction packets length ratio, bi-direction bytes, packets/duration (second), bytes/duration (second), mean packets arrived time (duration/packets), bi-direction packets ratio, bi-direction packets, mean packet length, and tos. Applications: FTP, HTTPS, HTTP, POP3, IMAP, SMTP, SQLnet, Oracle, DNS, NTP, LDAP, Kazaa, Bittorrent, Gnutella, eDonkey, Media Player, Real, SSH, klogin, Telnet, GAME Halflife, SIP, and Skype.	R:60-94, P:60-93, A:84-94, G-mean: 58.30-71.80

Neto et al. [152] represented traffic classes by using empirical distributions that correspond to the traffic classes signatures, aiming to develop a classifier based in the dark mechanism that combined both Kolmogorov-Smirnov and Chi-square tests. Chi-square was also used in [153] to test if a set of data follows a specific distribution with a degree of confidence.

3) SHANNON ENTROPY

Gomes et al. in [155] used entropy to emphasize and recognize VoIP P2P traffic flows that belonged to a VoIP session. The developed classifier aimed to identify the flow used in the conversation and focused on the specific characteristics of the voice codec instead of the application used in the VoIP session. In [154], Wang et al. used entropy to classify traffic more deeply. In [156], Zhou et al., used entropy for evaluation of encrypted traffic classification. In Table 12, there is a summary of works about the use of entropy.

V. DISCUSSION AND OPEN ISSUES

A. DISCUSSION

Distance and divergence computations are advanced methods of statistical analysis that can be used for classification and,

TABLE 12. Works related to other statistical methods (Correlation information, Kolmogorov-Smirnov, Chi-Square, Entropy).

Method	Work	Year	Characteristics	Performance
Correlation Information	Zhang *et al.* [149]	2012	Features: flows sharing, and period of time. Applications: DNS, P2P, SSH/SSL, and FTP. Technique: BoF model-based.	FM:20-99, A:90
	Dong *et al.* [150]	2012	Features: high port number, low port number, bytes of flow, packets of flow, the average packet payload length, the average packet length, the average packet header length, duration flow duration, the average packet arrival interval of flow, byte number per second, packets number ratio of bidirectional flow, packets number per second, packet length ratio of bidirectional flow and byte number ratio of bidirectional flow. Applications: unspecified.	N/A
	Zhang *et al.* [61]	2013	Features: volume of bytes, size and number of packets, inter-packet time, and number of flow statistical properties. Applications: SSL, SSH, and HTTP.	A:58-90, FM:60-95
	Zhang *et al.* [151]	2014	Features: client-to-server maximum packet bytes, number of packets, client-to-server average packet bytes, client-to-server minimum packet bytes, client-to-server minimum inter-packet time, the standard deviation of client-to-server packet bytes, server-to-client number of packets, server-to-client minimum packet bytes, and server-to-client maximum packet bytes.	A: 80-95, FM:88-95
	Aggarwal and Singh [135]	2017	Feature: flow statistical. Technique: discretized statistical.	A:65-95
Chi-Square test	Neto *et al.* [152]	2013	Features: length of the packets; Applications: HTTP, Skype, and P2P. Technique: sliding windows.	P:89.36-100, R:91.24-100
	Casino *et al.* [153]	2019	Feature: Chi-square Absolute value. Applications: Compression method ZIP, RAR, BZIP2, and GZIP.	A:68.68-94.72
Kolmogorov-Smirnov	Neto *et al.* [152]	2013	Features: length of the packets. Applications: HTTP, Skype, and P2P. Technique: sliding windows.	P:89.36-100, R:91.24-100
Entropy	Wang *et al.* [154]	2011	Features: frequencies of characters and entropy of consecutive bytes. Applications: encrypted files (AES, PGP and SSL) and compressed files (.gz, .rar, .zip), P2P torrent packets, torrent protocol, SMTP, and HTTP. Techniques: SVM and Sequential Forward Selection (SFS), KL, and JSD.	A: 69-81
	Gomes *et al.* [155]	2012	Features: length of the packets. Applications: Voip, SIP, and Skype.	S:78.57-100, Sp:99.51-100
	Zhou *et al.* [156]	2019	Features: packet's inter-arrival time, packet's sizes, and direction as the neural network's input. Applications: classes (Voip, Audio, browsing, chat, email, FTP, P2P, and video. Techniques: NN, SVM, Random Forest, Naive Bayes, and Logistical regression.	ROC:0.73-0.96, F1:33-95, P: 36-93, R:30-96.

in our context, were used for Internet traffic classification. Through the statistical properties, statistical traffic classification models may be created for a given application. For these methods, sometimes a learning phase is required to build a reference model that can be used to classify traffic.

Statistical classification, also known as logic based classification, allows traffic identification through statistical attributes of the flow. The packet length and duration, the traffic flow idle timing, and the time between packet arrivals are considered examples of statistical traffic attributes or measurements of flow level. On sight of traffic, statistical classification tends to assume and explore unique resources of each application, using data mining techniques to do so most of the time.

Statistical classifiers are light weight and do not require packet payload analysis. In addition, they can achieve the same precision as other methods found in the literature, even using fewer features. These advantages make them suitable candidates for the most restricted configurations. Also, given the current trend towards flow level monitors like NetFlow [157], the ability to operate on statistical characteristics only is an advantageous property for classifiers.

As for the computational complexity of statistical methods, Valenti et al. [7] show how tree-based statistical classification can sustain high rate of transference on off-the-shelf hardware.

Figure 5 shows the Network Visualization map created using the VosViewer tool. This map was created from the references cited in this article, and based on bibliographic data. The data was read from reference manager files .ris. We chose the co-authorship analysis with fractional counting method, that is the strength of the document is divided by the total number of authors. We do not ignore documents with a large number of authors. For the generation of our map, we chose at least 1 author per document and found 462 different authors and co-authors. For each author, the total number of co-authors was calculated and the authors with the greatest total link strength will be selected were selected for the chart.

B. OPEN ISSUES

In the literature, several significant types of research have been done on traffic classification and how to improve the performance of the classifiers, but there are still some challenges ahead. Considering the technologies and methods applied, most challenges still lie in classifying encrypted, unknown, and P2P traffic in real-time or timely with high precision and low processing power.

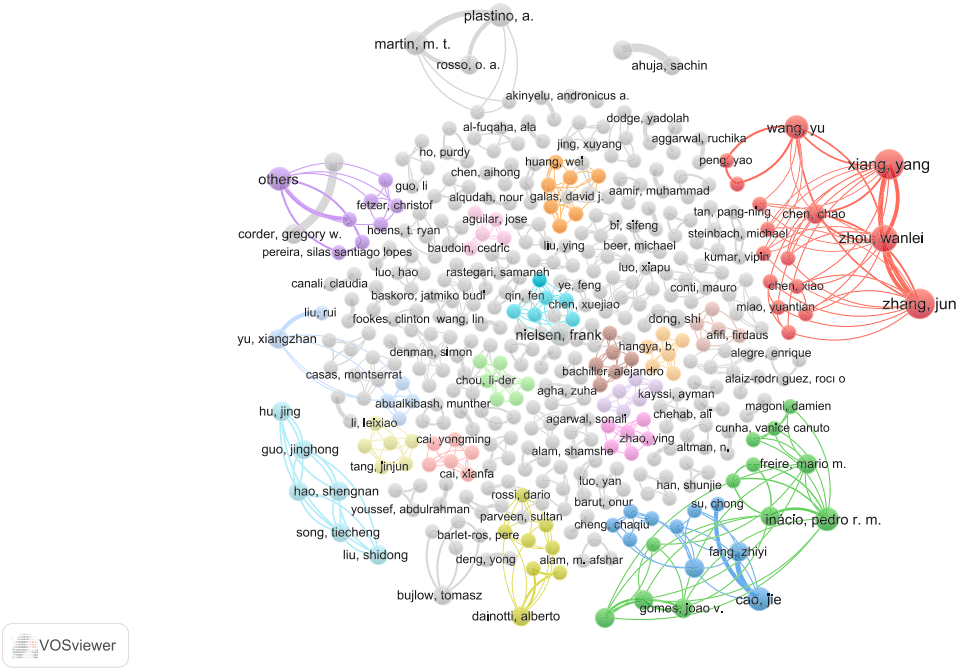

FIGURE 5. VosViewer network visualization map.

In this section, we outline some important open-ended research questions that need to be addressed in this field of research as follows:

- Although SVM has been widely used to classify traffic, traditional traffic classifiers based on SVM have their limitations, among them the high computational cost when it comes to memory, CPU, highly complex training and the difficulties to operate in real time, which makes the real time and timely classification unfeasible. Possible research directions may include the development of new SVM models to address the above issues, following the work in [137].
- SVM still faces resource selection imbalance issues in its training phase. For [158], solving the problems of imbalance in the SVM classification is kept an open issue.
- SVMs performance do not absolutely depend on the size of the training data, but on the quantity of Support Vectors (SVs). An open question for research is balancing data volume and complexity because according to [137], with the increase in training data, computational complexity and the occupation of computational resources will also grow significantly.
- One of the issues to be worked on when implementing an Internet traffic classifier using SVM, is choosing correctly the self parameter C because, according to [158], the classification is sensitive to C, in which, if not chosen correctly, SVM, even optimized, produces worse classification results.
- Explore the feasibility of the use of the Wootters Distance for encrypted Internet traffic classification, which, to the best of our knowledge, has not yet been investigated.
- Investigate the use of less explored statistical distances and divergences for encrypted Internet traffic classification, namely Bhattacharyya and Hellinger Distances and Jensen-Shannon Divergence. Although these statistical methods have been investigated for network security and intrusion detection, among other, as reported in this work, we did not find specific applications of these methods for classification of encrypted Internet traffic.
- Explore the combination of the SVM classifier with statistical divergences. In the literature, we find works that combine Euclidean Distance with the K-means algorithm for classifiers, and Kullback-Leibler combined with SVM. However, we did not find classifiers combined with Hellinger Distance, Wootters Distance, Jensen-Shannon Divergence, for example.

VI. CONCLUSION

The main purpose of this work was to explore statistical methods and techniques recently used or with the potential to be used in Internet network traffic classification. We provided an overview of the Internet traffic classification process as well as an insight into statistical methods with potential interest to be used as classifiers for encrypted Internet traffic, including those methods that have not yet been explored previously for Internet traffic classification. Then, we reviewed previously used statistical methods for Internet traffic classification, organized by distances, divergences, SVM, and other statistical methods. Through the literature review, we identified that the most used statistical method for traffic classification is the SVM method. In addition, we also identified several open issues that could be the subject of further research on this topic. More specifically, we identified statistical distances and divergences that have not been much explored regarding

to traffic classification. Actually, they could be used separately or combined with the SVM classifier in order to address challenging problems such as real-time traffic classification and encrypted traffic classification.

REFERENCES

[1] A. A. Mohamed, A. H. Osman, and A. Motwakel, "Classification of unknown internet traffic applications using multiple neural network algorithm," in *Proc. 2nd Int. Conf. Comput. Inf. Sci. (ICCIS)*, Oct. 2020, pp. 1–6.
[2] 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.
[3] J. Zhao, X. Jing, Z. Yan, and W. Pedrycz, "Network traffic classification for data fusion: A survey," *Inf. Fusion*, vol. 72, pp. 22–47, Aug. 2021.
[4] 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.
[5] Y. Peng, M. He, and Y. Wang, "A federated semi-supervised learning approach for network traffic classification," 2021, *arXiv:2107.03933*.
[6] R. Liu and X. Yu, "A survey on encrypted traffic identification," in *Proc. Int. Conf. Cyberspace Innov. Adv. Technol.*, Dec. 2020, pp. 159–163.
[7] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, "Reviewing traffic classification," in *Data Traffic Monitoring and Analysis* (Lecture Notes in Computer Science), vol. 7754, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Berlin, Germany: Springer, 2013, doi: 10.1007/978-3-642-36784-7_6.
[8] 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.
[9] X. Li, "Study on traffic flow base on RBF neural network," in *Proc. 6th Int. Conf. Measuring Technol. Mechatronics Autom.*, Jan. 2014, pp. 645–647.
[10] P. Mehta and R. Shah, "A survey of network based traffic classification methods," *Database Syst. J.*, vol. 7, no. 4, pp. 24–31, Jan. 2017.
[11] M. Cotton, L. Eggert, J. Touch, M. Westerlund, and S. Cheshire. *Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry*, RFC document 6335, pp. 1–33, Aug. 2011.
[12] K. L. Dias, M. A. Pongelupe, W. M. Caminhas, and L. D. Errico, "An innovative approach for real-time network traffic classification," *Comput. Netw.*, vol. 158, pp. 143–157, Jul. 2019.
[13] *WAN and Application Optimization Solution Guide, Cisco Validated Design*, document Version 1.1, Cisco Systems, Aug. 2008. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/nsite/enterprise/wan/wan_optimization/wan_opt_sg.pdf
[14] T. Bujlow, V. Carela-Español, and P. Barlet-Ros, "Independent comparison of popular DPI tools for traffic classification," *Comput. Netw.*, vol. 76, pp. 75–89, Jan. 2015.
[15] M. Lotfollahi, M. J. Siavoshani, R. S. H. Zade, and M. Saberian, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *Soft Comput.*, vol. 24, no. 3, pp. 1999–2012, 2020.
[16] 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. Surveys*, vol. 45, no. 3, pp. 1–40, Jun. 2013.
[17] Y. Liu, "A survey of machine learning based packet classification," in *Proc. Symp. Comput. Intell. Secur. Defence Appl. (CISDA)*, Jul. 2009, pp. 1–10.
[18] G. Sahoo and Y. Kumar, "Analysis of parametric & non parametric classifiers for classification technique using weka," *Int. J. Inf. Technol. Comput. Sci.*, vol. 4, no. 7, p. 43, Jul. 2012.
[19] S. Han, C. Qubo, and H. Meng, "Parameter selection in SVM with RBF kernel function," in *Proc. World Automat. Congr.*, Jun. 2012, pp. 1–4.
[20] Z. Zhang, J. T. Kwok, and D.-Y. Yeung, "Parametric distance metric learning with label information," in *Proc. IJCAI*, vol. 1450, Jan. 2003, pp. 1–20.
[21] H. Xu and Y. Deng, "Dependent evidence combination based on Shearman coefficient and Pearson coefficient," *IEEE Access*, vol. 6, pp. 11634–11640, 2017.
[22] M. Menéndez, J. Pardo, L. Pardo, and M. Pardo, "The Jensen–Shannon divergence," *J. Franklin Inst.*, vol. 334, no. 2, pp. 307–318, Mar. 1997.
[23] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. CT-15, no. 1, pp. 52–60, Feb. 1967.
[24] L. Su and H. White, "A nonparametric Hellinger metric test for conditional independence," *Econ. Theory*, vol. 24, no. 4, pp. 829–864, Apr. 2008.
[25] Y. Dodge and D. Commenges, *The Oxford Dictionary of Statistical Terms*. New York, NY, USA: Oxford Univ. Press Demand, Oct. 2006.
[26] J. Poza, C. Gómez, M. García, A. Bachiller, A. Fernández, and R. Hornero, "Analysis of spontaneous meg activity in mild cognitive impairment and alzheimer's disease using Jensen's divergence," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 1501–1504.
[27] F. J. Massey, Jr., "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951.
[28] R. Dunaytsev, D. Moltchanov, Y. Koucheryavy, O. Strandberg, and H. Flinck, "A survey of p2p traffic management approaches: Best practices and future directions," *Internet Eng.*, vol. 5, no. 1, pp. 318–330, Jun. 2012.
[29] A. Pradhan, "Support vector machine—A survey," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, Aug. 2012.
[30] A. Dainotti, A. Pescape, and K. C. Claffy, "Issues and future directions in traffic classification," *IEEE Netw.*, vol. 26, no. 1, pp. 35–40, Jan. 2012.
[31] E. A. Jamuna and S. V. Ewards, "Survey of traffic classification using machine learning," *Int. J. Adv. Res. Comput. Sci.*, vol. 4, no. 4, pp. 65–71, Apr. 2013.
[32] B. Li, J. Springer, G. Bebis, and M. H. Gunes, "A survey of network flow applications," *J. Netw. Comput. Appl.*, vol. 36, no. 2, pp. 567–581, Mar. 2013.
[33] S. Tavara, "Parallel computing of support vector machines: A survey," *ACM Comput. Surveys*, vol. 51, no. 6, pp. 1–38, Feb. 2019.
[34] T. Garrett, L. E. Setenareski, L. M. Peres, L. C. E. Bona, and E. P. Duarte, "Monitoring network neutrality: A survey on traffic differentiation detection," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2486–2517, 3rd Quart., 2018.
[35] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *Appl. Sci.*, vol. 9, no. 20, p. 4396, Oct. 2019.
[36] M. Shen, Y. Liu, L. Zhu, K. Xu, X. Du, and N. Guizani, "Optimizing feature selection for efficient encrypted traffic classification: A systematic approach," *IEEE Netw.*, vol. 34, no. 4, pp. 20–27, Jul. 2020.
[37] 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.
[38] P. Wang, X. Chen, F. Ye, and Z. Sun, "A survey of techniques for mobile service encrypted traffic classification using deep learning," *IEEE Access*, vol. 7, pp. 54024–54033, 2019.
[39] S. Alam, S. K. Sonbhadra, S. Agarwal, and P. Nagabhushan, "One-class support vector classifiers: A survey," *Knowl.-Based Syst.*, vol. 196, May 2020, Art. no. 105754.
[40] 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.
[41] J. Yan and J. Yuan, "A survey of traffic classification in software defined networks," in *Proc. IEEE 1st Int. Conf. Hot Inf.-Centric Netw. (HotICN)*, Aug. 2018, pp. 200–206.
[42] N. Alqudah and Q. Yaseen, "Machine learning for traffic analysis: A review," *Proc. Comput. Sci.*, vol. 170, pp. 911–916, Apr. 2020.
[43] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: A review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 857–900, Aug. 2019.
[44] S. Hussain, M. Abualkibash, and S. Tout, "A survey of traffic sign recognition systems based on convolutional neural networks," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, May 2018, pp. 0570–0573.
[45] O. Salman, I. H. Elhajj, A. Kayssi, and A. Chehab, "A review on machine learning–based approaches for internet traffic classification," *Ann. Telecommun.*, vol. 75, nos. 11–12, pp. 673–710, Jun. 2020.
[46] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering, version 2.3," Keele Univ., Keele, U.K., Durham Univ., Durham, U.K., Tech. Rep. EBSE 2007-001, 2007. [Online]. Available: https://www.researchgate.net/publication/258968007_Kitchenham_B_Guidelines_for_performing_Systematic_Literature_Reviews_in_software_engineering_EBSE_Technical_Report_EBSE-2007-01
[47] J. Yepes-Nuñez, G. Urrutia, M. Romero-Garcia, and S. Alonso-Fernandez, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Revista Espanola de Cardiologia*, vol. 74, no. 9, pp. 790–799, 2021.
[48] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Systematic Rev.*, vol. 10, no. 1, pp. 1–11, 2021.

[49] Sandvine. (May 2014). *The Global Internet Phenomena Report 2014*. [Online]. Available: https://www.sandvine.com/trends/global-internet-phenomena/
[50] M. Tamilkili, "A survey on recent traffic classification techniques using machine learning methods," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 12, pp. 368–373, Jun. 2013.
[51] M. Conti, L. V. Mancini, R. Spolaor, and N. V. Verde, "Analyzing Android encrypted network traffic to identify user actions," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 1, pp. 114–125, Jan. 2015.
[52] M. Shafiq, X. Yu, A. K. Bashir, H. N. Chaudhry, and D. Wang, "A machine learning approach for feature selection traffic classification using security analysis," *J. Supercomput.*, vol. 74, no. 10, pp. 4867–4892, 2018.
[53] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 1, pp. 5–16, Jan. 2007.
[54] S. E. Gómez, L. Hernández-Callejo, B. C. Martínez, and A. J. Sánchez-Esguevillas, "Exploratory study on class imbalance and solutions for network traffic classification," *Neurocomputing*, vol. 343, pp. 100–119, May 2019.
[55] H. Shi, H. Li, D. Zhang, C. Cheng, and W. Wu, "Efficient and robust feature extraction and selection for traffic classification," *Comput. Netw. Int. J. Comput. Telecommun. Netw.*, vol. 119, pp. 1–16, Jun. 2017.
[56] S. Angra and S. Ahuja, "Machine learning and its applications: A review," in *Proc. Int. Conf. Big Data Anal. Comput. Intell. (ICBDAC)*, Mar. 2017, pp. 57–60.
[57] D. Bzdok, N. Altman, and M. Krzywinski, "Points of significance: Statistics versus machine learning," *Nature Methods*, vol. 15, pp. 1–7, Apr. 2018.
[58] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Paper e0194889, doi: 10.1371/journal.pone.0194889.
[59] S. Rastegari, P. Hingston, and C.-P. Lam, "Evolving statistical rulesets for network intrusion detection," *Appl. Soft Comput.*, vol. 33, pp. 348–359, Aug. 2015.
[60] Z. Chen, Z. Liu, L. Peng, L. Wang, and L. Zhang, "A novel semi-supervised learning method for internet application identification," *Soft Comput.*, vol. 21, no. 8, pp. 1963–1975, Apr. 2017.
[61] J. Zhang, C. Chen, Y. Xiang, W. Zhou, and A. V. Vasilakos, "An effective network traffic classification method with unknown flow detection," *IEEE Trans. Netw. Service Manage.*, vol. 10, no. 2, pp. 133–147, Jun. 2013.
[62] N. Antunes and M. Vieira, "On the metrics for benchmarking vulnerability detection tools," in *Proc. 45th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, Jun. 2015, pp. 505–516.
[63] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika*, vol. 14, no. 4, p. 1502, 2016.
[64] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.
[65] G. W. Corder and D. I. Foreman, *Nonparametric statistics: A Step-By-Step Approach*. Hoboken, NJ, USA: Wiley, Apr. 2014.
[66] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander, "Parametric and nonparametric linkage analysis: A unified multipoint approach," *Amer. J. Hum. Genet.*, vol. 58, no. 6, p. 1347, Jun. 1996.
[67] S. Boslaugh, *Statistics in a Nutshell: A Desktop Quick Reference*. Sebastopol, CA, USA: O'Reilly Media, Nov. 2012.
[68] S. Glen. (Aug. 2021). *Correlation Coefficient: Simple Definition, Formula, Easy Steps*. Accessed: Aug. 3, 2020. [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
[69] N. Pandis, "The chi-square test," *Amer. J. Orthodontics Dentofacial Orthopedics*, vol. 150, no. 5, pp. 898–899, 2016.
[70] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[71] M. T. Martin, A. Plastino, and O. A. Rosso, "Statistical complexity and disequilibrium," *Phys. Lett. A*, vol. 311, nos. 2–3, pp. 126–132, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0375960103004912, doi: 10.1016/S0375-9601(03)00491-2.
[72] M. Basseville, "Divergence measures for statistical data processing—An annotated bibliography," *Signal Process.*, vol. 93, no. 4, pp. 621–633, Apr. 2013.
[73] M. T. Martin, A. Plastino, and O. A. Rosso, "Statistical complexity and disequilibrium," *Phys. Lett. A*, vol. 311, nos. 2–3, pp. 126–132, May 2003.
[74] C. A. Solà, "Recent statistical methods based on distances," *Contrib. Sci.*, vol. 2, pp. 183–192, Aug. 2002.
[75] M. Markatou and E. M. Sofikitou, "Statistical distances and the construction of evidence functions for model adequacy," *Frontiers Ecol. Evol.*, vol. 7, p. 447, Nov. 2019.
[76] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. London, U.K.: Pearson Education India, Mar. 2019.
[77] F. Nielsen, "On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid," *Entropy*, vol. 22, no. 2, p. 221, Feb. 2020.
[78] F. Nielsen, "On the Jensen–Shannon symmetrization of distances relying on abstract means," *Entropy*, vol. 21, no. 5, p. 485, 2019.
[79] F. C. Schweppe, "On the Bhattacharyya distance and the divergence between Gaussian processes," *Inf. Control*, vol. 11, no. 4, pp. 373–395, Oct. 1967.
[80] S. Bi, M. Broggi, and M. Beer, "The role of the Bhattacharyya distance in stochastic model updating," *Mech. Syst. Signal Process.*, vol. 117, pp. 437–452, Feb. 2019.
[81] D. J. Weller-Fahy, B. J. Borghetti, and A. A. Sodemann, "A survey of distance and similarity measures used within network intrusion anomaly detection," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 70–91, 1st Quart., 2015.
[82] S. P. Vaidya and P. V. S. S. R. C. Mouli, "Adaptive, robust and blind digital watermarking using Bhattacharyya distance and bit manipulation," *Multimedia Tools Appl.*, vol. 77, no. 5, pp. 5609–5635, Mar. 2018.
[83] A. Majtey, P. Lamberti, M. Martin, and A. Plastino, "Wootters' distance revisited: A new distinguishability criterium," *Eur. Phys. J. D-Atomic, Mol., Opt. Plasma Phys.*, vol. 32, no. 3, pp. 413–419, Jan. 2005.
[84] J. Wu and R. J. Karunamuni, "Efficient Hellinger distance estimates for semiparametric models," *J. Multivariate Anal.*, vol. 107, pp. 1–23, May 2012.
[85] C. Su and J. Cao, "Improving lazy decision tree for imbalanced classification by using skew-insensitive criteria," *Int. J. Speech Technol.*, vol. 49, no. 3, pp. 1127–1145, Mar. 2019.
[86] V. González-Castro, R. Alaiz-Rodríguez, and E. Alegre, "Class distribution estimation based on the Hellinger distance," *Inf. Sci.*, vol. 218, pp. 146–164, Jan. 2013.
[87] D. Cieslak, T. Hoens, N. Chawla, and W. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Mining Knowl. Discovery*, vol. 24, no. 1, pp. 136–158, 2012.
[88] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, Jul. 2006.
[89] C. Delpha, D. Diallo, and A. Youssef, "Kullback–Leibler divergence for fault estimation and isolation: Application to gamma distributed data," *Mech. Syst. Signal Process.*, vol. 93, pp. 118–135, Sep. 2017.
[90] D. Galas, G. Dewey, J. Kunert-Graf, and N. Sakhanenko, "Expansion of the Kullback–Leibler divergence, and a new class of information metrics," *Axioms*, vol. 6, no. 4, p. 8, Apr. 2017.
[91] S. Parveen, S. K. Singh, U. Singh, and D. Kumar, "A comparative study of traditional and Kullback–Leibler divergence of survival functions estimators for the parameter of Lindley distribution," *Austrian J. Statist.*, vol. 48, no. 5, pp. 45–53, Jul. 2019.
[92] P. J. Moreno, P. Ho, and N. Vasconcelos, "A Kullback–Leibler divergence based kernel for SVM classification in multimedia applications," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2003, pp. 1385–1392.
[93] D. Kvitsiani, S. Ranade, B. Hangya, H. Taniguchi, J. Z. Huang, and A. Kepecs, "Distinct behavioural and network correlates of two interneuron types in prefrontal cortex," *Nature*, vol. 498, no. 7454, pp. 363–366, May 2013.
[94] A. M. Kowalski, M. T. Martín, A. Plastino, O. A. Rosso, and M. Casas, "Distances in probability space and the statistical complexity setup," *Entropy*, vol. 13, no. 6, pp. 1055–1075, Jun. 2011.
[95] M. T. Martin, A. Plastino, and O. A. Rosso, "Generalized statistical complexity measures: Geometrical and analytical properties," *Phys. A, Stat. Mech. Appl.*, vol. 369, no. 2, pp. 439–462, Sep. 2006.
[96] W. Peng, A. Chen, and J. Chen, "Using general master equation for feature fusion," *Future Gener. Comput. Syst.*, vol. 82, pp. 119–126, May 2018.
[97] C. Cortes and V. Vapnik, "Support vector machines," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[98] T. Marwala, "Support vector machines," in *Handbook of Machine Learning*. Singapore: Wold Scientific, Dec. 2018, pp. 97–112.
[99] E. J. C. Suárez, "Tutorial sobre Máquinas de Vectores Soporte (SVM)," Dpto. de Inteligencia Artificial, ETS de Ingeniería Informática, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, 2016. [Online]. Available: https://www.cartagena99.com/recursos/alumnos/apuntes/Tema8._Maquinas_de_Vectores_Soporte.pdf

[100] C.-C. Liu, Y. Chang, C.-W. Tseng, Y.-T. Yang, M.-S. Lai, and L.-D. Chou, "SVM-based classification mechanism and its application in SDN networks," in *Proc. 10th Int. Conf. Commun. Softw. Netw. (ICCSN)*, Jul. 2018, pp. 45–49.
[101] S. E. N. Fernandes, A. L. Pilastri, L. A. M. Pereira, R. G. Pires, and J. P. Papa, "Learning kernels for support vector machines with polynomial powers of sigmoid," in *Proc. 27th SIBGRAPI Conf. Graph., Patterns Images*, Aug. 2014, pp. 259–265.
[102] M. Singla and K. Shukla, "Robust statistics-based support vector machine and its variants: A survey," *Neural Comput. Appl.*, vol. 32, pp. 1–22, Dec. 2019.
[103] S. S. L. Pereira, J. E. B. Maia, and J. L. De Castro e Silva, "ITCM: A real time internet traffic classifier monitor," 2015, *arXiv:1501.01321*.
[104] H. Singh, "Performance analysis of unsupervised machine learning techniques for network traffic classification," in *Proc. 5th ICACCT*, Feb. 2015, pp. 401–404.
[105] B. Schmidt, A. Al-Fuqaha, A. Gupta, and D. Kountanis, "Optimizing an artificial immune system algorithm in support of flow-based internet traffic classification," *Appl. Soft Comput.*, vol. 54, pp. 1–22, May 2017.
[106] P. Zhu, S. Zhang, H. Luo, and Z. Wu, "A semi-supervised method for classifying unknown protocols," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Mar. 2019, pp. 1246–1250.
[107] M. Zanin, "The reasonable effectiveness of data in ATM," in *Proc. SESAR Innov. Days*, Nov. 2013, pp. 1–5.
[108] C. Canali and R. Lancellotti, "Automatic virtual machine clustering based on Bhattacharyya distance for multi-cloud systems," in *Proc. Int. Workshop Multi-Cloud Appl. Federated Clouds (MultiCloud)*, Apr. 2013, pp. 45–52.
[109] M. A. Dinani, P. Ahmadi, and I. Gholampour, "Efficient feature extraction for highway traffic density classification," in *Proc. 9th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Nov. 2015, pp. 14–19.
[110] H. Sadreazami, A. Mohammadi, A. Asif, and K. N. Plataniotis, "Distributed-graph-based statistical approach for intrusion detection in cyber-physical systems," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 1, pp. 137–147, Mar. 2017.
[111] M. I. Sameen and B. Pradhan, "A two-stage optimization strategy for fuzzy object-based analysis using airborne LiDAR and high-resolution orthophotos for urban road extraction," *J. Sensors*, vol. 2017, pp. 1–17, Feb. 2017.
[112] J. B. Baskoro, A. Wibisono, and W. Jatmiko, "Bhattacharyya distance-based tracking: A vehicle counting application," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Oct. 2017, pp. 439–444.
[113] E. Laz, "Optimal cost allocation in centralized and decentralized detection systems using Bhattacharyya distance," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2017, pp. 1170–1173.
[114] M. H. Shah and X. Dang, "Novel feature selection method using Bhattacharyya distance for neural networks based automatic modulation classification," *IEEE Signal Process. Lett.*, vol. 27, pp. 106–110, 2020.
[115] C.-H. Liu, P. Pawelczak, and D. Cabric, "Primary user traffic classification in dynamic spectrum access networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 11, pp. 2237–2251, Nov. 2014.
[116] J. Safarik, M. Voznak, F. Rezac, and J. Slachta, "Application of artificial intelligence on classification of attacks in IP telephony," *Adv. Inf. Sci. Appl.*, vol. 2, pp. 373–378, Nov. 2014.
[117] Z. Luo, P.-M. Jodoin, S.-Z. Li, and S.-Z. Su, "Traffic analysis without motion features," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3290–3294.
[118] C. Wang, T. T. N. Miu, X. Luo, and J. Wang, "SkyShield: A sketch-based defense system against application layer DDoS attacks," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 559–573, Mar. 2018.
[119] A. Kumari and U. Thakar, "Hellinger distance based oversampling method to solve multi-class imbalance problem," in *Proc. 7th Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, Nov. 2017, pp. 137–141.
[120] Z. Liu, R. Wang, N. Japkowicz, Y. Cai, D. Tang, and X. Cai, "Mobile app traffic flow feature extraction and selection for improving classification robustness," *J. Netw. Comput. Appl.*, vol. 125, pp. 190–208, Jan. 2019.
[121] S. Liang, "Feature extraction of broken glass cracks in road traffic accident site based on deep learning," *Complexity*, vol. 2021, pp. 1–12, May 2021.
[122] J. Garcia and T. Korhonen, "Efficient distribution-derived features for high-speed encrypted flow classification," in *Proc. Workshop Netw. Meets AI ML (NetAI)*, 2018, pp. 21–27.
[123] M. Zareapoor, P. Shamsolmoali, and M. A. Alam, "Advance DDoS detection and mitigation technique for securing cloud," *Int. J. Comput. Sci. Eng.*, vol. 16, no. 3, pp. 303–310, May 2018.
[124] T. Zhi, Y. Liu, J. Wang, and H. Zhang, "Resist interest flooding attacks via entropy–SVM and Jensen–Shannon divergence in information-centric networking," *IEEE Syst. J.*, vol. 14, no. 2, pp. 1776–1787, Jun. 2020.
[125] O. Barut, R. Zhu, Y. Luo, and T. Zhang, "TLS encrypted application classification using machine learning with flow feature engineering," in *Proc. 10th Int. Conf. Commun. Netw. Secur.*, Nov. 2020, pp. 32–41.
[126] J. Chen, D. Wu, Y. Zhao, N. Sharma, M. Blumenstein, and S. Yu, "Fooling intrusion detection systems using adversarially autoencoder," *Digit. Commun. Netw.*, vol. 7, no. 3, pp. 453–460, Aug. 2021.
[127] J. Kim, J. Hwang, and K. Kim, "High-performance Internet traffic classification using a Markov model and Kullback–Leibler divergence," *Mobile Inf. Syst.*, vol. 2016, pp. 1–13, 2016.
[128] J. Xu, S. Denman, C. Fookes, and S. Sridharan, "Detecting rare events using Kullback–Leibler divergence: A weakly supervised approach," *Exp. Syst. Appl.*, vol. 54, pp. 13–28, Jul. 2016.
[129] X. Zhang, F. Qiu, and F. Qin, "Identification and mapping of winter wheat by integrating temporal change information and Kullback–Leibler divergence," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 76, pp. 26–39, Apr. 2019.
[130] V. C. Cunha, A. A. Zavala, P. R. Inácio, D. Magoni, and M. M. Freire, "Classification of encrypted internet traffic using Kullback–Leibler divergence and Euclidean distance," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.* Cham, Switzerland: Springer, Mar. 2020, pp. 883–897.
[131] S. Hao, J. Hu, S. Liu, T. Song, J. Guo, and S. Liu, "Network traffic classification based on improved DAG-SVM," in *Proc. Int. Conf. Commun., Manage. Telecommun. (ComManTel)*, Dec. 2015, pp. 256–261.
[132] S. Hao, J. Hu, S. Liu, T. Song, J. Guo, and S. Liu, "Improved SVM method for internet traffic classification based on feature weight learning," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Oct. 2015, pp. 102–106.
[133] Z. Fan and R. Liu, "Investigation of machine learning based network traffic classification," in *Proc. ISWCS*, Aug. 2017, pp. 1–6.
[134] J. Cao, Z. Fang, G. Qu, H. Sun, and D. Zhang, "An accurate traffic classification model based on support vector machines," *Int. J. Netw. Manage.*, vol. 27, no. 1, p. e1962, Jan. 2017.
[135] R. Aggarwal and N. Singh, "A new hybrid approach for network traffic classification using SVM and naïve Bayes algorithm," *Int. J. Comput. Sci. Mobile Comput*, vol. 6, pp. 168–174, Aug. 2017.
[136] Y. Miao, Z. Ruan, L. Pan, J. Zhang, and Y. Xiang, "Comprehensive analysis of network traffic data," *Concurrency Comput., Pract. Exper.*, vol. 30, no. 5, p. e4181, Mar. 2018.
[137] G. Sun, T. Chen, Y. Su, and C. Li, "Internet traffic classification based on incremental support vector machines," *Mobile Netw. Appl.*, vol. 23, no. 4, pp. 789–796, Aug. 2018.
[138] A. A. Akinyelu and A. E. Ezugwu, "Nature inspired instance selection techniques for support vector machine speed optimization," *IEEE Access*, vol. 7, pp. 154581–154599, 2019.
[139] J. Tang, X. Chen, Z. Hu, F. Zong, C. Han, and L. Li, "Traffic flow prediction based on combination of support vector machine and data denoising schemes," *Phys. A, Stat. Mech. Appl.*, vol. 534, Nov. 2019, Art. no. 120642.
[140] M. Aamir and S. M. Ali Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 4, pp. 436–446, May 2021.
[141] C. Luo, C. Huang, J. Cao, J. Lu, W. Huang, J. Guo, and Y. Wei, "Short-term traffic flow prediction based on least square support vector machine with hybrid optimization algorithm," *Neural Process. Lett.*, vol. 50, pp. 2305–2322, Mar. 2019.
[142] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Phys. A, Stat. Mech. Appl.*, vol. 517, pp. 29–35, Mar. 2019.
[143] A. Şentaş, İ. Tashiev, F. Küçükayvaz, S. Kul, S. Eken, A. Sayar, and Y. Becerikli, "Performance evaluation of support vector machine and convolutional neural network algorithms in real-time vehicle type and color classification," *Evol. Intell.*, vol. 13, no. 1, pp. 83–91, Mar. 2020.
[144] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," *Exp. Syst. Appl.*, vol. 176, Aug. 2021, Art. no. 114885.
[145] S. Sankaranarayanan and S. Mookherji, "SVM-based traffic data classification for secured iot-based road signaling system," in *Research Anthology on Artificial Intelligence Applications in Security*. Hershey, PA, USA: IGI Global, Nov. 2021, pp. 1003–1030.
[146] J. Cao and Z. Fang, "Network traffic classification using genetic algorithms based on support vector machine," *Int. J. Secur. Its Appl.*, vol. 10, no. 2, pp. 237–246, Feb. 2016.

[147] M. Sabzekar, M. H. Y. Moghaddam, and M. Naghibzadeh, "TCP traffic classification using relaxed constraints support vector machines," in *Integration of Practice-Oriented Knowledge Technology: Trends and Prospectives*, M. Fathi, Ed. Berlin, Germany: Springer, 2013, doi: 10.1007/978-3-642-34471-8_11.
[148] D. L. Quoc, V. D'Alessandro, B. Park, L. Romano, and C. Fetzer, "Scalable network traffic classification using distributed support vector machines," in *Proc. IEEE 8th Int. Conf. Cloud Comput.*, Jun. 2015, pp. 1008–1012.
[149] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, and Y. Guan, "Network traffic classification using correlation information," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 1, pp. 104–117, Jan. 2013.
[150] S. Dong, "Measure correlation analysis of network flow based on symmetric uncertainty," *KSII Trans. Internet Inf. Syst.*, vol. 6, no. 6, pp. 1649–1667, 2012.
[151] 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.
[152] M. Neto, J. V. Gomes, M. M. Freire, and P. R. M. Inacio, "Real-time traffic classification based on statistical tests for matching signatures with packet length distributions," in *Proc. 19th IEEE Workshop Local Metrop. Area Netw. (LANMAN)*, Apr. 2013, pp. 1–6.
[153] F. Casino, K.-K. R. Choo, and C. Patsakis, "HEDGE: Efficient traffic classification of encrypted and compressed packets," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 11, pp. 2916–2926, Nov. 2019.
[154] Y. Wang, Z. Zhang, L. Guo, and S. Li, "Using entropy to classify traffic more deeply," in *Proc. 6th Int. Conf. Netw., Archit., Storage*, Jul. 2011, pp. 45–52.
[155] J. V. Gomes, P. R. M. Inacio, M. Pereira, M. M. Freire, and P. P. Monteiro, "Identification of peer-to-peer VoIP sessions using entropy and codec properties," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 10, pp. 2004–2014, Oct. 2013.
[156] K. Zhou, W. Wang, C. Wu, and T. Hu, "Practical evaluation of encrypted traffic classification based on a combined method of entropy estimation and neural networks," *ETRI J.*, vol. 42, no. 3, pp. 311–323, Jan. 2020.
[157] A. Campazas-Vega, I. S. Crespo-Martínez, Á. M. Guerrero-Higueras, C. Álvarez-Aparicio, and V. Matellán, "Analysis of NetFlow features' importance in malicious network traffic detection," in *Proc. Comput. Intell. Secur. Inf. Syst. Conf.* Cham, Switzerland: Springer, Sep. 2021, pp. 52–61.
[158] S. Rezvani, X. Wang, and F. Pourpanah, "Intuitionistic fuzzy twin support vector machines," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 11, pp. 2140–2151, Nov. 2019.

VANICE CANUTO CUNHA received the master's degree in electrical engineering from the University of Brasília (UNB). She is currently pursuing the Ph.D. degree in computer engineering with the University of Beira Interior (UBI), Portugal, in cooperation with the Université de Bordeaux, Talence, France. She is also a Professor at the Institute of Computing (IC), Federal University of Mato Grosso (UFMT), Campus Cuiabá, Brazil.

ARTURO ZAVALA ZAVALA received the degree in statistics and informatics from the Universidad Nacional Agraria La Molina, Peru, in 1988, and the master's and Ph.D. degrees in statistics from the University of São Paulo (USP), Brazil, in 2001 and 2004, respectively. He is currently an Associate Professor IV at the Federal University of Mato Grosso. He has experience in the area of probability and statistics, with an emphasis on experimental planning, working mainly on the following topics: regional development, agribusiness and sustainability, biofuels, dea and multidimensional poverty.

DAMIEN MAGONI (Senior Member, IEEE) received the M.Eng. degree from Télécom Paris, and the M.Sc. and Ph.D. degrees from the University of Strasbourg, in 1999 and 2002, respectively. He has been a Visiting Researcher at various institutions around the world, including the AIST at Tsukuba, the University of Sydney, the University of Michigan at Ann Arbor, and University College Dublin. He has been a Full Professor of computer science at the University of Bordeaux, since 2008. From 2002 to 2008, he was an Associate Professor at the University of Strasbourg. Some of his research has been supported by grants from the European Union, the CNRS, and Science Foundation Ireland. He has co-published over 80 refereed research papers. He also has authored several open-source software for networking research and teaching. His latest contributions are the virtual network device and the network mobilizer, which jointly enable the emulation of mobile networks. His main research interests include computer communications and networking, with a focus on internet architecture, protocols, and applications. He has reviewed for over 20 academic journals and has been in the TPC of numerous high-level conferences. He is a Senior Member of the ACM.

PEDRO R. M. INÁCIO (Senior Member, IEEE) received the 5 year B.Sc. degree in mathematics and computer science and the Ph.D. degree in computer science and engineering from the University of Beira Interior (UBI), Portugal, in 2005 and 2009, respectively. The Ph.D. work was performed in the enterprise environment of Nokia Siemens Networks Portugal S.A., through a Ph.D. Grant from the Portuguese Foundation for Science and Technology. He has been a Professor of computer science with the UBI, since 2010, where he lectures on subjects related to information assurance and security, and computer-based simulation, for graduate and undergraduate courses, namely, the B.Sc., M.Sc., and Ph.D. courses in computer science and engineering. He is currently an Instructor of the UBI Cisco Academy. He is also a Researcher with the Instituto de Telecomunicações.

MÁRIO M. FREIRE (Member, IEEE) received the five-year B.Sc. degree in electrical engineering and the two-year M.Sc. degree in systems and automation from the University of Coimbra, Portugal, in 1992 and 1994, respectively, and the Ph.D. degree in electrical engineering and the Habilitation degree in computer science from the University of Beira Interior (UBI), Portugal, in 2000 and 2007, respectively. He is currently a Full Professor of computer science at the UBI, which he joined in the Fall of 1994. In April 1993, he did one-month internship at the Research Centre of Alcatel-SEL (now Nokia Networks), Stuttgart, Germany. He is the coauthor of seven international patents, a co-editor of eight books published in the Springer LNCS book series, and the coauthor of about 130 papers in international journals and conferences. His main research interests include computer systems and networks, including network and systems virtualization, cloud and edge computing, and security and privacy in computer systems and networks. He is a member of the IEEE Computer Society and of the Association for Computing Machinery. He serves as a member of the Editorial Board for the ACM SIGAPP Applied Computing Review, serves as an Associate Editor for the *Security and Privacy* journal (Wiley) and of the *International Journal of Communication Systems* (Wiley), and served as an Editor for IEEE Communications Surveys and Tutorials, during 2007–2011. He served as a technical program committee member for several IEEE international conferences and is the Co-Chair of the Track on Networking of ACM SAC 2021. He is a Chartered Engineer by the Portuguese Order of Engineers.