

Open-Source Framework for Encrypted Internet and Malicious Traffic Classification

Ofek Bader^{*†‡}, Adi Lichy^{*†‡}, Amit Dvir^{*†‡}, Ran Dubin^{*†‡}, Chen Hajaj^{†‡§}

^{*}Department of Computer Science, Ariel University, Israel

[‡]Ariel Cyber Innovation Center, Ariel University, Israel

[†]Data Science and Artificial Intelligence Research Center, Ariel University, Israel

[§]Department of Industrial Engineering & Management, Ariel University, Israel



Abstract—Internet traffic classification plays a key role in network visibility, Quality of Services (QoS), intrusion detection, Quality of Experience (QoE) and traffic-trend analyses. In order to improve privacy, integrity, confidentiality, and protocol obfuscation, the current traffic is based on encryption protocols, e.g., SSL/TLS. With the increased use of Machine-Learning (ML) and Deep-Learning (DL) models in the literature, comparison between different models and methods has become cumbersome and difficult due to a lack of a standardized framework. In this paper, we propose an open-source framework, named OSF-EIMTC, which can provide the full pipeline of the learning process. From the well-known datasets to extracting new and well-known features, it provides implementations of well-known ML and DL models (from the traffic classification literature) as well as evaluations. Such a framework can facilitate research in traffic classification domains, so that it will be more repeatable, reproducible, easier to execute, and will allow a more accurate comparison of well-known and novel features and models. As part of our framework evaluation, we demonstrate a variety of cases where the framework can be of use, utilizing multiple datasets, models, and feature sets. We show analyses of publicly available datasets and invite the community to participate in our open challenges using the OSF-EIMTC.

Index Terms—Framework, Encrypted Traffic, Machine learning, Software

1 INTRODUCTION

Internet traffic classification works tackle the internet traffic classification problem from different approaches (e.g., payload-based and behavior-based) and categorize the data representation methods in various ways. For instance, statistical features are taken from the network flows. Classical machine-learning models have been shown to be applicable in the scope of internet traffic classification [1]–[4]. Several works recently used Natural Language Processing (NLP) techniques such as transforming the flow into a language to use word embedding [5].

Recently, there has been a huge change in the internet traffic where new network protocols such as QUIC, HTTP/2, HTTP/3 and new privacy concerned protocols such as TLS 1.3 and DoH [6]–[10] have been introduced. Consequently, the ability of the well-known solutions, which are based on DPI, ML, and DL classification systems, will be affected and will require extensive research.

Regarding any network classification problem based on the ML/DL pipeline, the researchers must cope with several important questions: beginning with which dataset should be used, how to extract a set of features, how to construct a new model, and how to compare the new model with the well-known models. Due to the lack of a common shared framework for internet traffic classification, the above tasks are difficult and tiresome. For example, in the application network traffic classification task [11]–[15], researchers would like to compare their features and models using the same framework, especially in systems using the same subset of features (e.g., 784 payload bytes of the flow) [12], [13], [16], [17] and minor changes in the models such as a similar CNN architecture [11], [18].

Therefore, the contribution of this work is as follows: we introduce a novel Open-Source Framework for Encrypted Internet and Malicious Traffic Classification (OSF-EIMTC). The OSF-EIMTC provides a full ML/DL pipeline as can be depicted in Figure 1. The pipeline can be characterized by the following phases: dataset selection, feature extraction, model selection, and evaluation. In the dataset selection phase, the framework allows access to well-known datasets (e.g., ISCXVPN2016, Ariel). In the feature extraction phase, the framework can extract sets of state-of-the-art features (e.g., flow and packet payload bytes [12], [13], [16]–[18]), with the ability to add new feature sets (via plugins). In the model selection phase, the framework also allows access to standard ML classifiers (e.g., RF, SVM) and provides implementations of state-of-the-art deep-learning classification models (e.g., MalDIST [16]). In the evaluation phase, the framework implements standard model evaluation metrics (e.g., accuracy, recall). The framework can be for researchers as a benchmarking platform to evaluate their approaches by comprehensive to well known features and solutions.

- *The authors want to thank Antonio Montieri for his help in the implementation of the original DISTILLER system and also want to thank Tal Shapira for his help in the implementation of the FlowPic feature extraction and its respective model. This work was supported by the Ariel Cyber Innovation Center in conjunction with the Israel National Cyber Directorate in the Prime Minister's Office.*

Furthermore, researchers can extend the framework platform with new features and models in order to contribute to the research community. We utilize the framework to evaluate different models over different datasets, showcasing the framework's advantages. We also used the framework to analyze several public state-of-the-art datasets and providing several interesting insights about the datasets.

The rest of this paper is organized as follows. Section 2 describes the related works. Section 3 discusses the framework implementation. Section 4 presents various datasets and their analysis. Section 5 presents runs of the framework evaluation. Section 6 presents online challenges, and in conclusion, section 7 presents a summary and discusses future work.

2 RELATED WORK

In recent years, deep-learning models have become the prominent method for network traffic classification. Some works that utilize deep-learning such as [12], [18], [19] even utilize raw data of the payload of the flow to feed deep-learning algorithms, thereby demonstrating that hand-crafting features are not always needed.

The works that utilize deep-learning span over multiple scopes and domains, with works that focus on the classification of the operating system, browser, and application levels [1]; mobile app identification [2], [3]; and even webpage fingerprinting [4]. Several works converted the network flow into an image to harness image processing techniques and equivalent Deep-Learning (DL) architectures [2], [11], [17], [18], [20], [21]. Newer works incorporated the Ordinary Differential Equation Network (ODENet) within a DL architecture to classify uni-directional network flows [22]. If we shift the focus to the cyber domain, many works tackle the task of malware network traffic detection and classification. [17]–[19], [23]–[30].

Along with detection or classification methods, the data in use have great importance as well. Some publicly available datasets such as ISCXPVPN2016 are widely used in the literature in works such as [11]–[13], [16], [22], [23], [31]. Where the works adopt the dataset for the purpose of classifying a network flow according to its encapsulation type (tunneled via VPN or not), traffic type (e.g., video/audio/chat), and application (e.g., Skype/Netflix). However, the dataset contains extensive background noise [32], such as unrelated BlueStacks [13] and Dropbox [16] network traffic. Researchers may need to clean and preprocess the data before using it. Therefore, access to several clean datasets in one place is an advantage for researchers. For example, Barut et al., [32] compiled their dataset with well-known datasets for both malware detection and application type categorization.

As many works attempt to solve similar tasks, researchers may want to compare their own models and features to other proposed traffic classification methods as part of the evaluation process. NetML [23] is an example of a framework that provides its own compiled data files from publicly available sources with features already extracted and ready to use. In both normal and malicious network traffic tasks, researchers need to implement their own models and compare them to others while using the same

provided feature files. The features were extracted with an accelerated feature extraction library by Intel. However, access to the library and the code used for feature extraction purposes are not given or shared. This complicates the ability to replicate the same features and methods with another dataset. Moreover, NetML does not implement any state-of-the-art deep-learning (DL) models. nPrint [33] is another tool that unifies the representation of each packet into a standard presentation that is amenable for representation learning. Though nPrint is designated to automate the process of machine learning pipelines, it does not propose an option for the extraction of custom features. While nPrint is integrated with an AutoML library (AutoGluon-Tabular), it does not support user-defined models, such as Deep-Learning neural networks (e.g., M1CNN [12], MalDIST [16]).

3 THE FRAMEWORK

Our framework is an open-source [34] that enables a comparison of multiple new and well-known features and state-of-the-art models for both network and malware traffic classification. The proposed framework allows the researcher to acquire a complete ML/DL pipeline with minimal time and effort as a benchmark to compare their new ideas and as a platform to plugin new features and models.

The envisioned flow of the ML/DL pipeline is presented in Figure 1, demonstrating the ability of the framework to create a complete pipeline scheme. The workflow comprises a total of 5 parts:

- 1) Data: The network traffic data files match the task that the researcher is attempting to solve, whether it be VPN detection, application network classification, or malicious traffic identification. The data is one of the most crucial components in research and experiments to achieve satisfactory results in real-world scenarios. **Our framework provides access to some of the well-known datasets e.g., [35]–[38] in an organized manner.** Some details about the well-known data sources can be found in Section 3.1 while their analysis can be found in Section 4 and in the Appendix.
- 2) Feature extraction: Many of the previous works used different tools for feature extraction, where some of them are not publicly available [23]. The lack of public access to the feature extraction tools affects the reproducibility of the same features. Furthermore, the existence of different tools complicates the process of extracting features, and makes it difficult to combine different feature sets from different tools. Moreover, currently combining new features with old ones is a complicated task. **The framework can extract various well-known features, for example, flow-based and packet size-related features (e.g., min, max, and mean of packet sizes of the flow). The framework can also extract TLS-related features such as TLS record size and direction as in [39]–[41]. Additionally, the framework is able to extract the well-known full features set of well-known works such as DeepMAL [17] payload bytes per packet, FlowPic**

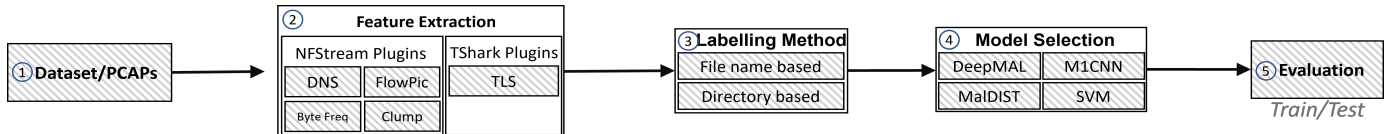


Fig. 1: The Open-Source Framework for Encrypted Internet and Malicious Traffic Classification Pipeline

images [11], and the features for the modalities of DISTILLER [13] and MalDIST [16]. Given the ability to extract standalone features (e.g., TLS features) and full feature sets such as FlowPic, facilitates the ability to create new feature sets that are a combination of well-known standalone features, and to add new features. More details on feature extraction are provided in Section 3.2.

- 3) Labelling method: every sample should have a respective label. Every feature set extracted from the sample should have the same label. The labelling part takes the extracted features for each sample and merges them with their respective label. By providing a rule or a naming scheme, it is possible to automatically associate any feature set created from flows/packets (or any other network sample) with their labels. The respective labelling can be done in several ways such as correlating them with their original files or directories. **The framework allows one to configure the labeling method with the suitable naming scheme.** See Section 3.3 for examples.
- 4) Model selection: Many models exist such as classical machine-learning and deep-learning neural networks. A comparison should be conducted for any new model (which can be based on older models). Therefore, **the framework provides classical ML/DL models and implementations of state-of-the-art deep-learning architectures such as MalDIST [16] for malware detection & classification and M1CNN [12] for application classification.** The models are discussed in detail in Section 3.4.
- 5) Evaluation: To determine the performance and predict the robustness and effectiveness of a model, along with its provided features, on a particular task requires accurate evaluation methods. Comparison of a variety of models is crucial for quality research work. **In order to make the process of a new evaluation or model comparison easier, common evaluation metrics (such as accuracy, recall, precision, and F1-score) are provided in the framework.**

In the upcoming sections, we elaborate on the core parts of the framework pipeline, which include datasets, feature extraction, labeling methods, and model selection.

3.1 Datasets

The framework provides a catalog of datasets for easy access. In this section, we describe the contents of selected datasets. We analyze some of the datasets in section 4.

3.1.1 USTC2016 [38]

This dataset, which can be found here in [38], contains 10 malware families and 10 types of benign traffic.

Dataset	Domain	Contents	Ref.
USTC2016	Malware, Apps	10 types of malware families and 10 consumer applications.	[16]–[18]
Stratosphere	Malware	Variety of captured and/or simulated malware traffic, some captures are mixed (benign and malware) and the rest are purely benign.	[16], [23], [32]
MTA	Malware	Variety of malware network samples.	[16], [26]
ISCX2016	Apps	Samples of applications of different categories, along with VPN encapsulated traffic.	[12], [13], [16], [22], [23], [31], [32]
Ariel (BOA2016)	Browsers, OS, Apps	Samples of web applications traffics labeled by browser and operating system.	[15], [42]
MAppGraph 2021	Mobile Apps	Samples of over 80 popular mobile applications in Google Play.	[43]

- The malware classes are: Cridex (Dridex), Geodo (Emotet), Htbot, Miuref, Neris, Nsis-a, Shifu, Tinba, Virut, Zeus.
- The benign classes are: BitTorrent, Facetime, FTP, Gmail, MySQL, Outlook, Skype, SMB, Weibo, WorldOfWarcraft.

Some works that used this dataset can be found in [16]–[18]

3.1.2 StratosphereIPS [37]

Is a dataset comprising of three parts: benign, malware, and mixed traffic. The dataset was generated by Stratosphere Laboratory, a part of the CTU University of Prague in the Czech Republic. Additional information such as a description of the behavior captured facilitates the labeling process. Some works that adopt this dataset can be found in [23], [40], [44]

3.1.3 MTA (Malware Traffic Analysis) [36]

The data source is a website (blog) that includes many types of malware infection traffic for analysis. The website contains many types of malware such as ransomware and exploit kits. As of 2013 to date, the blog is updated daily with relevant malware traffic, continuously adding new samples to the dataset. Using Intrusion-Detection Systems (IDS) and Antivirus software, every binary file in the PCAPs has been confirmed as malicious. Papers such as [16], [19], [26] used this dataset for malware detection.

3.1.4 ISCX2016 (ISCXVPN2016) [35]

This dataset consists of 150 PCAP files of different types of traffic and applications. Each PCAP file has an application category (e.g., Spotify, Facebook, YouTube, etc.), a traffic type category (e.g., streaming, VoIP, chat, etc.) and an encapsulation label (non-VPN/VPN). Some works that adopted this dataset can be found in [11]–[13], [16], [22], [23], [32].

Note that the data contains extensive noise, such as unrelated BlueStacks traffic, Dropbox traffic, and background traffic. Works such as [11] extracted only the relevant flows from each PCAP file, whether the audio flow of Skype or the video flow of Netflix. Many works dropped browser traffic files due to lack of sufficient accurate labeling, and p2p traffic such as BitTorrent due to lack of counterparts in non-encapsulated traffic.

3.1.5 Ariel (BOA2016) [15]

This dataset is from a paper in which the authors collected the data over a period of more than two months, in their lab, using a selenium web crawler for browser traffic. The dataset contains applications’ traffic such as YouTube, and Facebook, which are labeled as browser traffic, and Dropbox and TeamViewer that are labeled as non-browser traffic. The dataset contains more than 20,000 sessions. The average duration of a session was 518 seconds where on average each session had 520 forward packets (the average forward traffic size was 261K bytes) and 637 backward packets (the average backward traffic size was 615K bytes). Almost all of the flows are TLS encrypted. Examples of works that used this dataset include [15], [42].

3.1.6 MAppGraph dataset (2021) [43]

The authors collected traffic of 101 **mobile applications**, which are popular in Google Play. For each application, more than 30 hours of traffic were collected, resulting in more than 600 GB of traffic stored in PCAP files. However, the authors published a portion of the dataset, which amounts to 81 applications that weighs around 500 GB. Some of the applications include but not limited to: Facebook, Twitch, Instagram, Skype, Spotify, Google Meet, Soundcloud, and Zoom.

3.2 Feature extraction

Features are one of the key components in the success of a machine learning model. There are many scopes of features; one can extract features from a packet (e.g., size and time), protocol/header (e.g., type and information), uni-directional flow (e.g., number of uploaded or downloaded bytes), bi-directional flow (e.g., average packet size), or a time window. The feature extraction of the framework is used to extract features from packets/flow/session. Hence, we can also extract features from uni and bi-directional flows. In this section, we present some standalone features and features sets that the framework supports. Note that, the framework offers the flexibility of adding any arbitrary feature or feature set that can be defined and harvested from the sample.

Our feature extraction component is based on the NFStream package [45]. NFStream is a Python framework

that provides a fast, flexible, and expressive data structures designed to make working with PCAP files both easy and intuitive. Some of the features of nDPI, which is an open and extensible deep packet inspection library, are integrated into NFStream and provides additional flow information such as fingerprints, request server name, and application category detection. To further enrich the of types of features extracted, we utilize TShark [46] as a post-NFStream tool to extract TLS related features per flow.

We developed new plugins to extract state-of-the-art features and their preprocessing phase, to make it easier to compare different models and extensions to allow new research directions with ease.

As the feature selection is provided in ‘plugins’, one can select a set of plugins with various features sets to extract and assemble the required final feature data file. The users can extend the component by developing their own plugins. We envision that more researchers would publish their code using these plugins to make the experiments more easily reproducible and therefore facilitate their use in different scenarios with far greater ease.

Plugin	Description	Features	Ref.
ASN_info	Autonomous System	ASN number, country code and description	[24], [47]
DNS	Domain Name System	#IP addr in response, TTL, domain name text statistics.	[24], [25], [48]
n_bytes	Payload bytes of the first N bytes of the flow.	array of length of N with the value of each byte.	[12], [13], [16], [18]
ByteFrequency	Byte distribution of the first N packets.	array of length of 256 with the distribution of each byte.	[25], [49]
SmallPacketRatio	The ratio of small packets from all the packets	Decimal ratio (single number)	[26], [50]
Protocol Headers	IAT, size, direction TCP-win size of packets	A $(n, 4)$ matrix, 4 features per packet of n packets.	[13], [14], [16]
Clumps	Statistical features over consecutive packets in the same direction.	Min, max, mean, std-dev of clump lengths, sizes, and IAT.	[41], [49]

TABLE 1: A subset of plugins provided in the framework used for feature extraction.

State-of-the-art features plugins: While the framework has numerous different plugins, we only discuss a subset of the plugins developed as part of the feature extraction component.

- ASN_info (Autonomous system information) [24], [47]: This plugin extracts and attaches ASN related information such as ASN number, ASN country code and ASN description to each flow. For example, the following ASN information will be attached to a

biflow with the source IP address 131.202.240.87 and the destination IP address 178.237.19.228:

[*num, code, description*]

[**num=611, code=CA, description=NB-PEI-EDUCATION-COMPUTER-NETWORK - University of Toronto**] for the source IP and [**num=47764, code=RU, description=MAILRU-AS Mail.Ru**] for the destination.

Note that the information is pulled from a local file-based DB, which may need to be downloaded and updated from time to time.

- DNS [24], [25], [48]: Malicious actors have utilized Command & Control (C2) communication channels over the Domain Name Service (DNS) and, in some cases, have even used the protocol to exfiltrate data. Malicious actors have also infiltrated malicious data/payloads to the victim's system over the DNS. Malware uses DNS as a way to bypass the block of the C2 server's IP address when it becomes known that the IP address is compromised or is used for malicious intents. The DNS along with a domain name for the C2 server allows the malware to communicate with its C2 server when the malware changes its IP address. Some works that tackled the detection of encrypted malware traffic, such as [25], utilized features from DNS traffic such as the number of IP addresses returned by a DNS response, the amount of digits in a DNS response, TTL values, and the domain name length. The plugin extracts a wide variety of DNS-related features, such as the number of answers per type, Time-To-Live values, and character-based counters in the answer section, for example digit count, hyphen count, dot count, and more. Other works such as [48] used DNS parameters to try to tackle the classification of users behind NAT, as part of internet traffic classification.
- Early raw byte payload (*n_bytes*) [12], [13], [16], [18]: This plugin extracts the first *n* bytes of the flow's payload, which can span over multiple packets in both upload and download links. Works such as [16], [18] used the raw data of the flow to detect and classify malware families by their network traffic, while [12], [13] used it for encrypted network traffic classification tasks.
- Byte Frequency [25], [49]: This plugin extracts the frequency of each byte value ([0-255]) that appears in the payload of the first *N* packets.
- Small Packet Ratio [26], [50]: A statistical feature of

$$\frac{\#small_packets}{\#packets}$$

where *#small_packets* is the number of packets with small payload ($\leq threshold$). The small packet ratio was used by Hung and Sun [50] for the detection system of botnets and it was presented in [26].

- Protocol Headers [13], [14], [16]: As proposed for example in [14], this plugin extracts size, IAT (Inter-arrival time), direction and TCP win-size from the first *n* packets of the flow to form a (*n*, 4) sized matrix. Lopez-Martin et al. proposed to use $n = 20$

while [13] used $n = 32$. A lower number of packets may be used for time critical tasks such as detecting malicious traffic. Lopez-Martin et al. [14] analyzed the performance of a specific deep-learning model with different number of packets given during training, for the task of differentiating IoT services, and showed that there is no significant impact on the model's performance even with as little as 6 packets. In another paper, Rezaei et al. [51] used the 6 first packets of the flow to identify mobile applications, the first packets in a TLS flow are plain and contain meaningful information.

- SubFlows/Clumps [41], [49]: This plugin extracts clump/subflow related features for each direction. It aggregates the packets of a session into clumps (or subflows), where a clump/subflow is a group of one or more consecutive packets with the same direction. It then extracts data about those clumps such as the min, max, and mean of clump's sizes (number of packets in a clump) and clump's lengths (number of bytes), and internal clump data such as min, max, mean of packets lengths, along with inter-arrival time and more. The rationale behind such grouping is that the application traffic is scattered among several packets as part of the process of TCP segmentation. Mohammadreza et al. [41] used the clumping method in order to lower the dimensionality of data. In contrast, Hong-Yen et al. [49] named such clumps subflows and claimed that their characteristics can enable machine-learning models to more accurately distinguish different types of sections in a flow while trying to detect changing points of switching applications over VPN encrypted traffic.
- TLS features [41], [52]: As the use of encryption in network traffic has become popular in recent years, TLS based features can be beneficial for the classification of encrypted flows. The TLS protocol is built on top of a reliable transport protocol such as TCP or QUIC. The data is transferred in the form of TLS records, where one of the first records in the flow is the "Client Hello" type from the client to the server, to which then the server responds with a "Server Hello" record of its own; these records are not encrypted (without eSNI, ECH and up until TLS 1.3) and contain a vast configuration data of the flow, such as cipher suites, the application layer protocol and many more. Some works such as [41] and [52] used the type and number of cipher suits offered by the client in order to detect TLS traffic initiated by malware, as different TLS clients use or support a different default or predefined set of cipher suits. The information inside the non-encrypted records such as the list of available compression methods, the list of available TLS extensions, and the list of available cipher suits, can be used to generate a client fingerprint. One of the methods for such fingerprinting is JA3 [53]. Other works such as [15] used the SNI field to classify and label network flows of services/applications such as Twitter and Youtube, where [23] extracted 14 different TLS features to classify network attacks, applications, and

traffic types. Extraction of TLS features such as TLS record sizes, types and TLS clumps are embedded in the framework by utilizing TShark as a step inside the feature extraction process.

3.3 Implementation of Common Labeling Methods

The proposed framework allows the use of two well-known name based labeling methods, a file name based and a directory based labeling inspired by the TensorFlow's [54] built-in utility function of loading an image dataset from a directory [55].

- 1) **File name based labeling method:** public datasets such as ISCX2016 use filename based labeling, such as a PCAP file, where the file name "vpn_facebook_audio2.pcap" has the following labels: *encapsulation type = vpn, traffic type = audio, application = facebook*. An example of a file/directory structure is as follows:

```

data
├── vpn_facebook_audio2.pcap
├── skype_video2b.pcap
└── vpn_bittorrent.pcap

```

- 2) **Directory based labeling:** inspired by the TensorFlow's utility function for loading an image dataset (tf.keras.utils.image_dataset_from_directory) [55]. PCAP files can be arranged into directories, where the directory name determines the label of all files in it. This is useful when the data for each label is too large to be contained in a single file or spans across many different files. An example of a file/directory structure is:

```

data
├── audio
│   ├── vpn_facebook_audio2.pcap
│   └── voipbuster_4b.pcap
├── video
│   └── skype_video2b.pcap
└── p2p
    └── vpn_bittorrent.pcap

```

3.4 State-Of-The-Art Models

While the users can utilize the extracted features data for training with any classical machine learning model available (e.g., K-Nearest-Neighbors (KNN), Support-Vector-Machine (SVM) and Random Forest (RF)), we also provide state-of-the-art deep-learning models [11]–[13], [16]–[18].

- DeepMAL [17]: Marin et al. proposed two variants for malware network traffic detection, one is packet-based and the other is flow-based. The packet-based variant trains and predicts the data extracted from single packets, while the flow-based variant requires multiple packets from the same flow. The DeepMAL raw flows model (flow-based variant) is trained on an input of n **payload** bytes per packet for the first m packets of the session. the input size is (m, n) per instance. the plugin extracts (m, n) features per session as a matrix, where each byte is represented as a decimal value in $[0, 255]$.
- M2CNN [18]: Wang et al. proposed a custom LeNet-5 neural-network architecture for malware network traffic detection and classification by leveraging the ability of 2D convolution layers to recognize different image-like patterns by feeding the model named by us as M2CNN, with a matrix-shaped input of bytes (an 8-bit gray image). M2CNN is trained on the n first **payload** bytes of the flow. M2CNN takes the bytes as a matrix of size (\sqrt{n}, \sqrt{n}) . The n_bytes plugin extracts the first n **payload** bytes of the session (n features), where each byte is represented as a decimal value in $[0, 255]$.
- M1CNN [12]: Wang et al. proposed viewing the bytes not as an image, but as a sequential time series, so the authors represented the bytes as a single array of length of 784 and utilized 1D convolutional layers instead of 2D, aiming to tackle the task of normal network traffic classification. M1CNN is trained on the n first **payload** bytes of the flow, the M1CNN takes the bytes as a single array of size $(1, n)$. the n_bytes plugin extracts the first n **payload** bytes of the session (n features), where each byte is represented as a decimal value in $[0, 255]$.
- FlowPic [11]: Shapira and Shavitt used a custom LeNet-5 neural network architecture. A FlowPic is a 2D-histogram-image of IP packet sizes and relative time-of-arrival. The plugin creates a FlowPic for each defined time window in the session and saves it on the file system as a compressed NumPy file (.npz), if the session has multiple time windows, multiple FlowPics can be created.
- DISTILLER [13]: Aceto et al. proposed a multi-task multi-modal deep-learning model used for classification of network traffic flows. the input to the model is of size 912, where the first 784 features are the 784 payload bytes of the flow, and the next 128 features are the features proposed by Lopez-Martin et al. [14] of [Size, IAT, direction, TCP win-size] of the first 32 packets.
- MalDIST [16]: In [16], the authors proposed a DISTILLER based variant that is used for malware detection and classification. The variant has a third novel modal of shape $(5, 14)$ inspired by the STNN model [56], resulting in a total number of 982 features. In the third modal, the packets in each flow are categorized into 5 categories, where then 14 statistical features are extracted from each category. The features of each category are then aligned in a row, resulting in an image with a shape of $(5, 14)$.
- A Custom DISTILLER Variant: One may wish to build a customized DISTILLER variant with new and/or different modalities. Since the model is a multitask model, i.e., predicts multiple classifications for the same flow, it is also possible to configure the number of tasks that the model is aimed for.

For each of the state-of-the-art models, we can use the model as standalone and the framework also provides the plugins that extracts their required features from raw PCAP files along with **the necessary preprocessing steps** that are required to perform before feeding it into the model for

training and predicting.

4 FRAMEWORK DATASETS ANALYSIS

In this section, we focus on some state-of-the-art publicly available datasets. We analyze each dataset with some insightful characteristics and bring conclusions. The datasets vary in their domains, from malware traffic to VPN-tunneled traffic.

We decided to present here one of the dataset analysis (USTC2016), while other three can be found in the Appendix. The main reason for using the framework to analyze the datasets is to shed light on the disadvantages of the well-known datasets, allowing the scientific community take these insights into account when using them. Moreover, we encourage the community to publish new updated, and diverse datasets.

The statistics and characteristics of the USTC dataset are presented in Figure 2. Fig 2b shows that the average number of packets per unidirectional flow in classes such as BitTorrent, Facetime, Outlook, and Skype is 1 (where Facetime has no downlink packets at all), which results in the inability to calculate flow duration (Fig 2d) because at least 2 packets are required for such calculations. Similarly in Fig 2e, the Inter-Arrival Time statistic couldn't be calculated from many flows that belong to benign classes. The high disparity between the value distribution of benign and malware flows in these figures of time-related features is evident. One can utilize these statistical features in models to differentiate between benign and malware traffic. Furthermore, the protocol distribution in fig 2f demonstrates that the benign traffic has quite a low number of different classes that use the UDP protocol, whereas the rest use TCP. Other protocols such as ICMPv6, ICMP, and IGMP can only be found in malware traffic, although in low quantities. Figure 2g shows that compared to benign flows, there is a high number of the malware flows that didn't successfully complete the three-way handshake that is required in the TCP protocol before data can be exchanged between the two endpoints, with one exception of Tinba, which contain UDP-based flows almost exclusively.

4.1 Datasets Disadvantages

Any dataset that is utilized for research has drastic effects on the results and conclusions that arise from it, and some issues exist with the publicly available datasets in the wild. The importance of the dataset becomes more highlighted when considering that other components in the ML/DL pipeline depend on it, such as features that are extracted from the data.

- **Quality:** The low quality of the provided network capture files (PCAPs), such as the USTC-TFC2016 dataset, among all the modifications to it such as anonymization, cleaning, and preprocessing that can degrade the data more, may result in inaccurate, non-representative samples of real-life scenarios. While the dataset contains network traffic captures of interesting applications such as MySQL and Facetime, evaluating a model on this dataset may not fully indicate the true prediction performance of it.

- **Format:** Some datasets do not contain raw traffic capture files (PCAPs), but rather a set of features already extracted [42] or truncated binary representation of each flow [57], which requires special care and non-standard processing. This becomes a problem when trying to compile a dataset from multiple sources. Without the raw files, it is impossible to extract new features with standard tools.
- **Date:** Datasets are quite old, for instance, the IS-CXVPN2016, USTC-TFC2016 and Ariel (BOA2016) datasets are from 2016. Still, these datasets are the default datasets used in most state-of-the-art works [11]–[13], [16], [22], [23], [31]¹. Applications change over time, along with the increasing trend of adopting encryption with almost every network traffic flow. For example, the network traffic of Skype or Netflix in 2016, might behave differently than the network traffic of Skype or Netflix in 2021.
- **Scale:** There is a bit of disconnection between the research in academia to actual implementation and usage of such innovations in the industry. The datasets feature a lower variety of classes compared to the much bigger scale required in industrialized solutions that sometimes make use of thousands and tens of thousands of different classes. The lack of a sufficient quantity of different classes makes it impossible to evaluate models and features in scenarios that employ a large set of classes in multiple metrics such as the model's performance and training/predicting time.

5 FRAMEWORK EVALUATION

In this section, we demonstrate how the framework can be used in multiple scenarios: **I.** We evaluated our framework in the naive case of replicating the results of previous work (Shapira and Shavitt FlowPic [11]). **II.** We used a set of three different models (i.e., M1CNN, MalDIST, and RF) on two different datasets (ISCX2016 [35], and Ariel [15]) to demonstrate the framework's ability to easily transfer the same experimental design (model and features) to a new dataset. **III.** We provide an example of how one can enhance the feature set of a model, using the plugins provided in our framework. Our main goal, across all three scenarios, is to show that our framework allows one to choose the dataset, the features sets, and the models to train, while replicating previous works, or transferring previous works to new datasets and/or new feature sets and models. The code that was used in this section is publicly available at [58].

As stated above, we began by replicating previous work using our framework. We ran the FlowPic's model on the ISCX2016 [35] dataset (using the same flows that the authors used in their evaluation) to validate the implementation of both its feature extraction and the model as provided in the framework. We began by executing the learning pipeline, from data to evaluation, that the framework offers. This

¹ Other datasets that are used are either lab recordings that are not publicly available, or datasets that are too small for adequate evaluation.

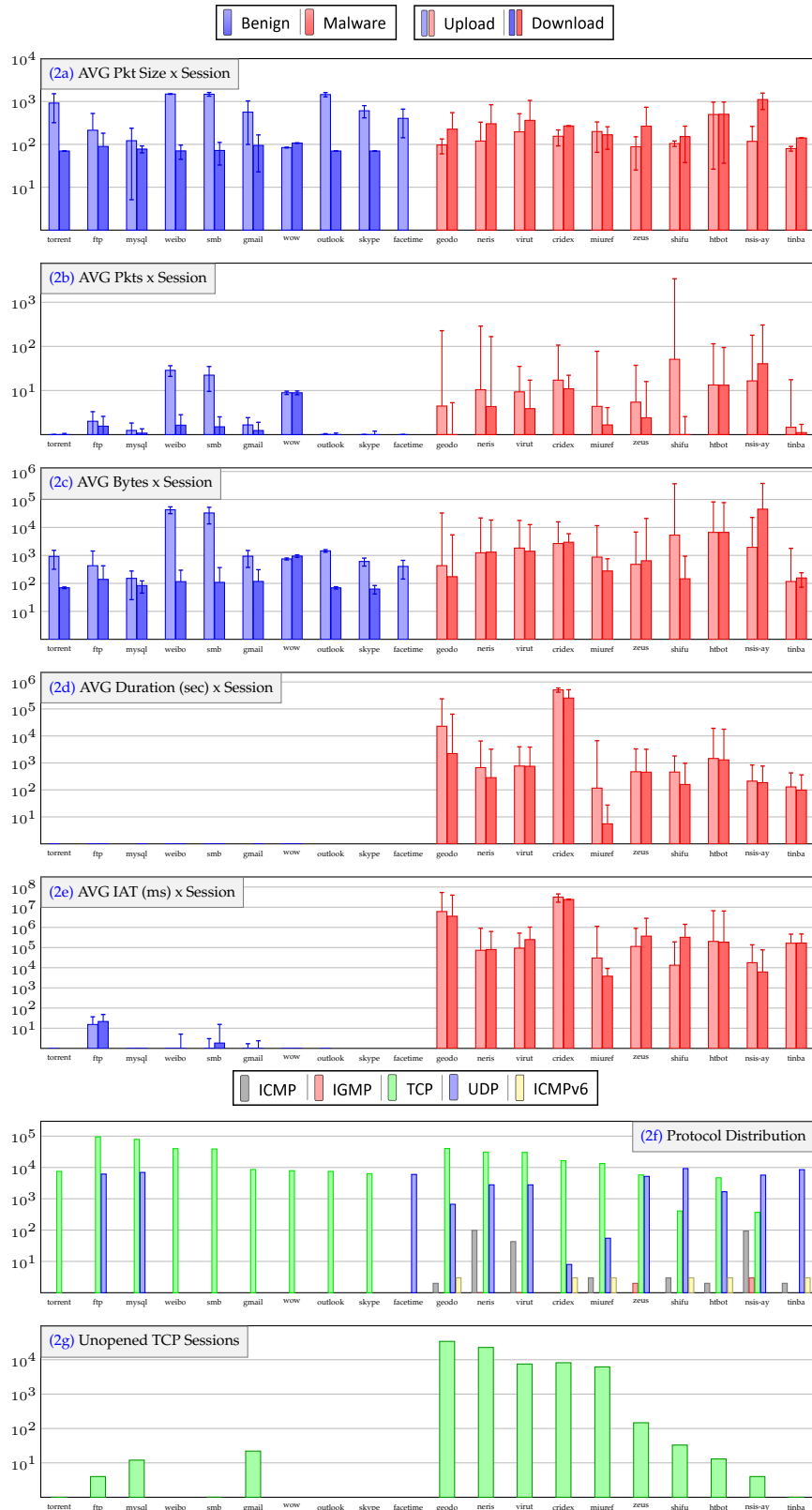


Fig. 2: Analysis of USTC2016.

includes, choosing the dataset, extracting FlowPic images as features from the dataset, selecting the matching deep-learning model, and then evaluating it. Similar to the results reported in [11], the model achieved over a 95% accuracy

score with our test set on the task of categorizing flows to their traffic type (audio/video/chat and so) over VPN or non-VPN traffic.

We then tested another framework functionality by eval-

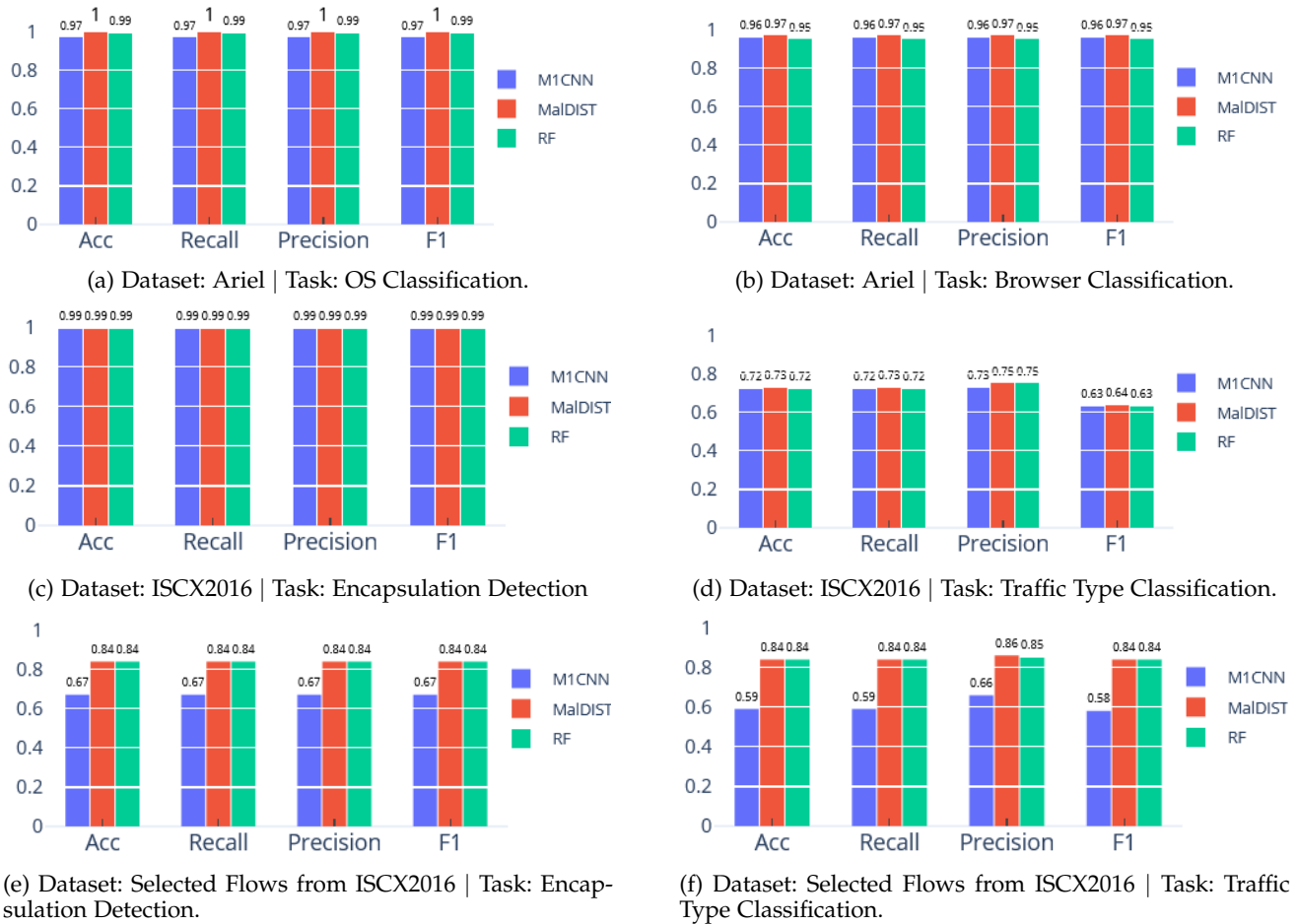


Fig. 3: Framework Evaluation - Same Model on Different Dataset

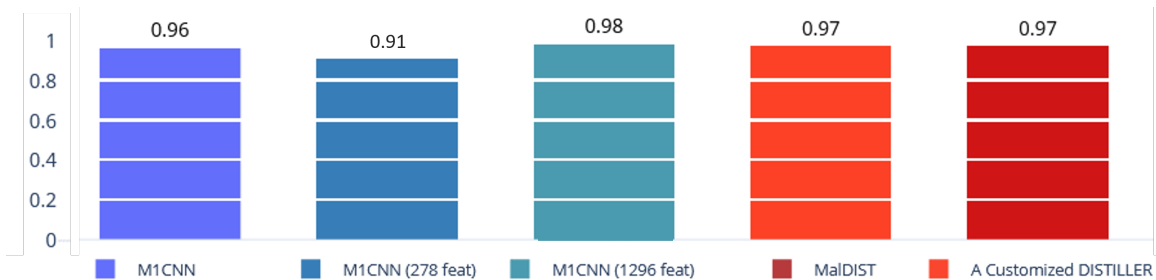


Fig. 4: Framework Evaluation - Feature Engineering on the Ariel Dataset for Browser Classification (The Score Values are for all Metrics of Accuracy, Recall, Precision and F1).

uating a set of models and their respective features on different datasets. The datasets were ISCX2016 [35] and Ariel [15]. We used the raw datasets without any preprocessing steps before extracting the features to show the applicability of the framework. We also included a third dataset based on selected flows from the ISCX2016 dataset, which were shared with us by the authors of FlowPic [11], for highly accurate flow-based labeling. From each dataset, we extracted the first 784 payload bytes (using the *n_bytes* plugin), Lopez-Martin's features of Protocol Headers fields (using the *Protocol Headers* plugin), STNN-inspired features (using the *STNN* plugin), and packet & TLS record clumps features. In light of the fact that both datasets already use filename-

based labeling of the data, we utilized the labeling method based on filenames (which is included in the framework). We selected M1CNN, MalDIST, and Random Forest (RF) as our models for evaluation. For the random forest model, we selected 24 clump-based features of both packets and TLS records. The results are presented in Fig 3. Using the Ariel dataset, the models performed superbly with a score of over 95% for all metrics in both OS and browser classification tasks, perfectly replicating the original results as reported in [15]. On the other hand, the models demonstrated several difficulties classifying flows by their encapsulation (VPN or non-VPN) and traffic type (e.g., video, audio, chat, or file transfer) with the ISCX2016 dataset. In this run, the

framework alleviated the process of evaluating models over different datasets. While the deep-learning models, such as M1CNN and MalDIST, were fed their proposed features (extracted using their respective plugins), we carefully selected 24 features for the Random Forest classifier.

One of the main parts of the learning pipeline is the feature engineering phase. Therefore, for the last scenario, we demonstrate the framework plugins ability, by feeding the deep-learning architecture with different input sizes and features. For M1CNN, we used an input size of 278 features, where the first 200 features were as suggested by the authors (i.e., the first 200 payload bytes of the flow out of the suggested 784 payload bytes), 70 features of a flattened array of STNN-inspired features, and a total of 8 other crafted statistical features (generated by the *PacketRelativeTime*, *SmallPacketPayloadRatio*, and *ResReqDiffTime* plugins).

We repeated the experiment of the second scenario with several changes. We fed the M1CNN model with 1296 features comprising the 784 payload bytes as suggested by the authors, along with an additional 512 features of byte frequency in the first 6 packets (using the *NPacketsByteFrequency* plugin), namely 256 features for each direction of the flow (byte values are in [0,255]).

WLOG, we evaluated the model on the Ariel dataset, choosing the browser classification task. The M1CNN (278 feat) model achieved a result of 91% for all metrics as depicted in Figure 4, which is a bit lower than the results yielded by M1CNN with the default input size and features with which we experimented as detailed in the previous paragraph (i.e., 96% for all metrics, see Fig 3). Figure 4 demonstrate that in the case of M1CNN (1296 feat), the evaluation of the model resulted in 98% for all metrics, surpassing all others.

Intrigued by the results of this M1CNN (1296 feat), we decided to use the DISTILLER architecture with customized plug-in components, in terms of modalities (inputs) and tasks (outputs). Therefore, we built a variant of DISTILLER which contained two modalities and two tasks (OS and browser). The first modality took the first 784 payload bytes of a flow as input, while the second modality took the byte frequencies (the same features that were fed into our last M1CNN) as input. This custom DISTILLER variant achieved a 97% score for all metrics for the task of browser classification. as can be seen in Fig. Fig 4. The figure shows that our framework provides the ability to use new features over well-known models. Note that in some cases this can improve the results (e.g., M1CNN - 1296 feat), while in other cases (e.g., M1CNN - 278 feat) it may decrease them. We encourage researchers to experiment with the feature engineering ability of the framework, which is simple and easy, while being aware that not every transformation is beneficial at the end of the road.

6 ONLINE CHALLENGES

In this section, we present online challenges hosted by EvalAI [59], [60]. EvalAI is an open-source platform for evaluating and comparing machine-learning models at scale. We created three challenges in the malware traffic detection and classification domain.

The three challenges are as follows:

- 1) Detection of malware traffic (binary): Distinguish between benign and malware traffic.
- 2) Classification of malware traffic (multi-class): Classify known malware family traffic.
- 3) Zero-day detection (binary): Detect unknown malware family traffic.

The public datasets that we utilized include:

- Benign: ISCX2016 [35], StratosphereIPS [37], and the benign subset of USTC2016 [38].
- Malware: MTA [36] and the malware subset of USTC2016 [38].

For each challenge, we provide the PCAP files for the *train* and *test* sets. Due to the class imbalance in the test sets of the first two challenges, we decided to use F1-score as the leading metric that the leader-board will follow to determine the best results. For the third challenge, zero-day detection, we decided to use the metrics of the detection rate, indicated by the True Positive Rate (TPR) and the False Alarm Rate (FAR), which is also known as the false positive rate. The leading metric in this challenge is a combination of the two:

$$TPR \cdot (1 - FAR) \quad (1)$$

Where the TPR and FAR values are calculated by the equations of:

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ FAR &= \frac{FP}{TN + FP} \end{aligned} \quad (2)$$

The description of the detailed challenges can be found on the challenges page on GitHub: <https://github.com/ArielCyber/OSF-EIMTC-Challenge>. We encourage the participants to use our framework [34] while attempting to tackle the presented challenges.

7 SUMMARY

The problem of classifying encrypted traffic will become more popular and important once the adoption of privacy-concerned and encrypted protocols such as DoH and QUIC will increase and gain momentum. This calls for extensive research and collaboration by developing and extending the framework with full ML/DL pipelines to tackle new protocols and other tasks in various domains. As a result, this paper presents a full pipeline framework. The framework creates a soft landing for new researchers to enter the domain of traffic classification. By providing organized access to datasets, feature extraction, and implementations for state-of-the-art deep-learning models, The framework allows to researchers to easily compare many models and feature sets, generating a rich comparison of a variety of solutions.

Researchers can also extend our framework to support new scopes of features, such as time windows and host-based features. By contributing to the framework, with new plugins of feature sets or implementations of newly proposed models, the research community will be able to save tremendous time when evaluating processes and

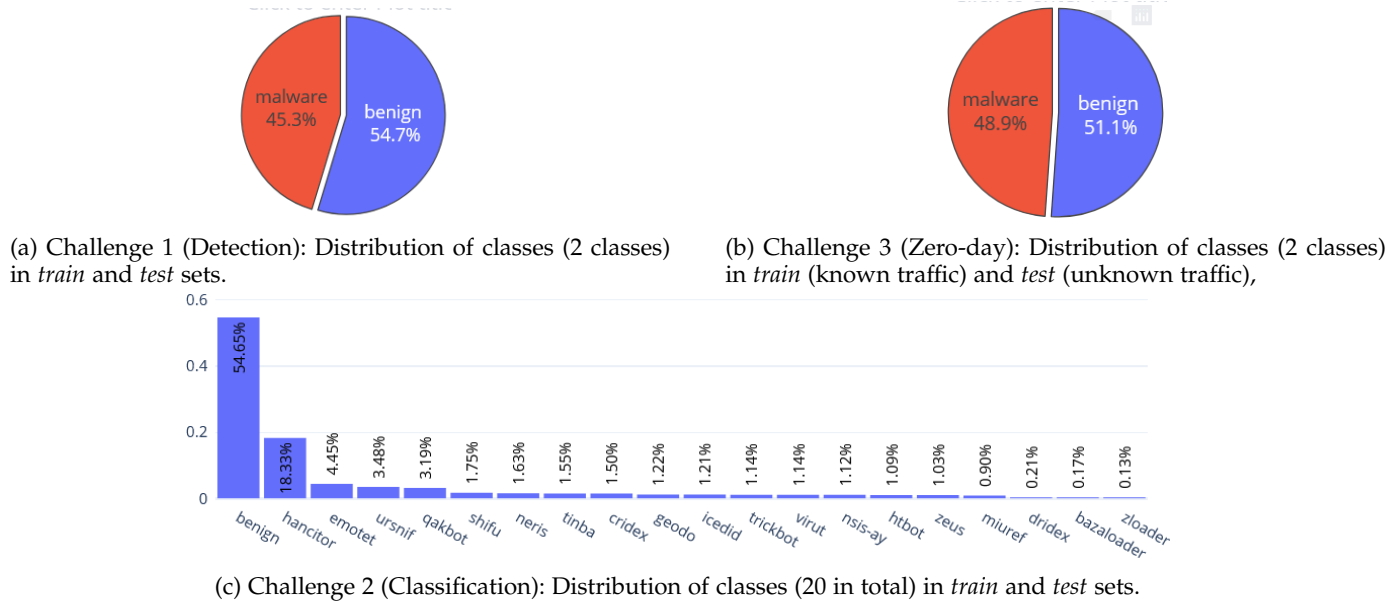


Fig. 5: Sample Distribution per Class in the Challenges.

models, facilitate more accurate comparisons, and enable reproducible experiments. New quality datasets with a vast number of classes and standardized formats (e.g., PCAP and PCAPNG), especially for new and upcoming protocols such as QUIC and TLS 1.3. are required to boost the research quality and the quality of the evaluations of proposed solutions. Any researcher can use the same plugins and tools that are provided in the proposed framework, to easily make use of new quality datasets. We expect that the online challenges that we have described in section 6 will promote the collaboration of the research community by contributing to the framework.

REFERENCES

- [1] J. Muehlstein, Y. Zion, M. Bahumi, I. Kirshenboim, R. Dubin, A. Dvir, and O. Pele, "Analyzing HTTPS encrypted traffic to identify user's operating system, browser and application," in *CCNC, Las Vegas, NV, USA, January 8-11, 2017*, pp. 1-6.
- [2] 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. 54 024-54 033, 2019.
- [3] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Robust smartphone app identification via encrypted network traffic analysis," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 63-78, Jan 2018.
- [4] M. Shen, Y. Liu, S. Chen, L. Zhu, and Y. Zhang, "Webpage fingerprinting using only packet length information," in *ICC, Shanghai, China, May 20-24, 2019*, pp. 1-6.
- [5] A. Dvir, A. K. Mamerides, R. Dubin, N. Golan, and C. Hajaj, "Encrypted video traffic clustering demystified," *Comput. Secur.*, vol. 96, p. 101917, 2020.
- [6] P. E. Hoffman and P. McManus, "DNS Queries over HTTPS (DoH)," RFC 8484, Oct. 2018. [Online]. Available: <https://rfc-editor.org/rfc/rfc8484.txt>
- [7] E. Rescorla, K. Oku, N. Sullivan, and C. A. Wood, "TLS Encrypted Client Hello," Internet Engineering Task Force, Internet-Draft draft-ietf-tls-esni-13, Aug. 2021, work in Progress. [Online]. Available: <https://datatracker.ietf.org/doc/html/draft-ietf-tls-esni-13>
- [8] M. Bishop, "Hypertext Transfer Protocol Version 3 (HTTP/3)," Internet Engineering Task Force, Internet-Draft draft-ietf-quic-http-34, Feb. 2021, work in Progress. [Online]. Available: <https://datatracker.ietf.org/doc/html/draft-ietf-quic-http-34>
- [9] Z. Chai, A. Ghafari, and A. Houmansadr, "On the importance of Encrypted-SNI (ESNI) to censorship circumvention," in *9th USENIX Workshop on Free and Open Communications on the Internet (FOCI 19)*. Santa Clara, CA: USENIX Association, Aug. 2019. [Online]. Available: <https://www.usenix.org/conference/foci19/presentation/chai>
- [10] E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.3," Internet Engineering Task Force, Internet-Draft draft-ietf-tls-rfc8446bis-03, Oct. 2021, work in Progress. [Online]. Available: <https://datatracker.ietf.org/doc/html/draft-ietf-tls-rfc8446bis-03>
- [11] T. Shapira and Y. Shavitt, "Flowpic: A generic representation for encrypted traffic classification and applications identification," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 2, pp. 1218-1232, 2021.
- [12] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks," in *ISI, Beijing, China, July 22-24*. IEEE, 2017, pp. 43-48.
- [13] G. Aceto, D. Ciunzo, A. Montieri, and A. Pescapè, "DISTILLER: encrypted traffic classification via multimodal multitask deep learning," *J. Netw. Comput. Appl.*, vol. 183-184, p. 102985, 2021.
- [14] M. L. Martín, B. Carro, A. Sánchez-Esguevillas, and J. Lloret, "Network traffic classifier with convolutional and recurrent neural networks for internet of things," 2017.
- [15] R. Dubin, A. Dvir, O. Pele, J. Muehlstein, Y. Zion, M. Bahumi, and I. Kirshenboim, "Analyzing https encrypted traffic to identify user's operating system, browser and application," in *IEEE Consumer Communications and Networking Conference*. IEEE, Jun. 2017.
- [16] O. Bader, A. Lichy, C. Hajaj, R. Dubin, and A. Dvir, "Maldist: From encrypted traffic classification to malware traffic detection and classification," in *Consumer Communications & Networking Conference (CCNC), IEEE Annual*. IEEE, 2022.
- [17] G. Marín, P. Casas, and G. Capdehourat, "Deepmal - deep learning models for malware traffic detection and classification," *CoRR*, vol. abs/2003.04079, 2020.
- [18] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *ICOIN, 2017*, pp. 712-717.
- [19] D. Kim, J. Han, J. Lee, H. Roh, and W. Lee, "Poster: Feasibility of malware traffic analysis through tls-encrypted flow visualization," in *ICNP 2020, Madrid, Spain, October 13-16*. IEEE, 2020, pp. 1-2.
- [20] O. Salman, I. H. Elhaji, A. I. Kayssi, and A. Chehab, "Data representation for CNN based internet traffic classification: a comparative study," *Multim. Tools Appl.*, vol. 80, no. 11, pp. 16 951-16 977, 2021.
- [21] S. Rezaei and X. Liu, "Deep learning for encrypted traffic clas-

- sification: An overview," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 76–81, 2019.
- [22] S. Roy, T. Shapira, and Y. Shavitt, "Fast and lean encrypted internet traffic classification," *Computer Communications*, vol. 186, 2022, pp. 166–173, 2022.
- [23] O. Barut, Y. Luo, T. Zhang, W. Li, and P. Li, "Netml: A challenge for network traffic analytics," *CoRR*, vol. abs/2004.13006, 2020.
- [24] D. Bekerman, B. Shapira, L. Rokach, and A. Bar, "Unknown malware detection using network traffic classification," in *2015 IEEE Conference on Communications and Network Security, CNS 2015, Florence, Italy, September 28-30, 2015*. IEEE, 2015, pp. 134–142.
- [25] B. Anderson and D. A. McGrew, "Identifying encrypted malware traffic with contextual flow data," D. M. Freeman, A. Mitrokotsa, and A. Sinha, Eds., 2016.
- [26] I. Letteri, G. D. Penna, L. D. Vita, and M. T. Grifa, "Mta-kdd'19: A dataset for malware traffic detection," in *CEUR Workshop Proceedings, Italy, February 4-7, M. Loreti and L. Spalazzi, Eds.*, vol. 2597, 2020, pp. 153–165.
- [27] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto, and F. Roli, "Yes, machine learning can be more secure! a case study on android malware detection," *IEEE Transactions on Dependable and Secure Computing*, 2017.
- [28] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *J. Netw. Comput. Appl.*, vol. 153, p. 102526, 2020.
- [29] A. Shabtai, L. Tenenboim-Chekina, D. Mimran, L. Rokach, B. Shapira, and Y. Elovici, "Mobile malware detection through analysis of deviations in application network behavior," *Comput. Secur.*, vol. 43, pp. 1–18, 2014.
- [30] J. G. de la Puerta, I. Pastor-López, B. Sanz, and P. G. Bringas, "Network traffic analysis for android malware detection," in *HAIS*, ser. Lecture Notes in Computer Science, vol. 11734, 2019, pp. 468–479.
- [31] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," in *ICISSP, Rome, Italy, February 19-21, O. Camp, S. Furnell, and P. Mori, Eds.*, 2016, pp. 407–414.
- [32] O. Barut, Y. Luo, T. Zhang, W. Li, and P. Li, "Multi-task hierarchical learning based network traffic analytics," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.
- [33] J. Holland, P. Schmitt, N. Feamster, and P. Mittal, "New directions in automated traffic analysis," in *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds.* ACM, 2021.
- [34] O. Bader, A. Lichy, C. Hajaj, R. Dubin, and A. Dvir, "OSF-EIMTC on GitHub." [Online]. Available: <https://github.com/ArielCyber/OSF-EIMTC>
- [35] "Canadian institute for cybersecurity iscxvpn2016," 2021. [Online]. Available: <https://www.unb.ca/cic/datasets/vpn.html>
- [36] B. Duncan, "Malware traffic analysis," 2021. [Online]. Available: <https://www.malware-traffic-analysis.net/>
- [37] Stratosphere, "Stratosphere laboratory datasets," 2015, retrieved March 13, 2020, from <https://www.stratosphereips.org/datasets-overview>.
- [38] W. Wang and D. Lu, "Ustc-tfc2016." [Online]. Available: <https://github.com/yungshenglu/USTC-TFC2016>
- [39] P. Brissaud, J. François, I. Chrisment, T. Cholez, and O. Bettan, "Encrypted HTTP/2 traffic monitoring: Standing the test of time and space," in *12th IEEE International Workshop on Information Forensics and Security, WIFS 2020, New York City, NY, USA, December 6-11, 2020*. IEEE, 2020, pp. 1–6.
- [40] M. J. D. Lucia and C. Cotton, "Detection of encrypted malicious network traffic using machine learning," in *MILCOM, Norfolk, VA, USA, November 12-14, 2019*, pp. 1–6.
- [41] M. MontazeriShatoori, L. Davidson, G. Kaur, and A. H. Lashkari, "Detection of doh tunnels using time-series classification of encrypted traffic."
- [42] S. Rezaei and X. Liu, "How to achieve high classification accuracy with just a few labels: A semi-supervised approach using sampled packets," *CoRR*, 2018.
- [43] T.-D. Pham, T.-L. Ho, T. Truong-Huu, T.-D. Cao, and H.-L. Truong, "MAppGraph: Mobile-App Classification on Encrypted Network Traffic using Deep Graph Convolution Neural Networks," in *Annual Computer Security Applications Conference (ACSAC 2021), Virtual Conference, December 2021*.
- [44] R. Moussaileb, N. Cuppens, J. Lanet, and H. L. Boudier, "Ransomware network traffic analysis for pre-encryption alert," in *FPS, Toulouse, France, November 5-7*, ser. Lecture Notes in Computer Science, vol. 12056. Springer, 2019, pp. 20–38.
- [45] "Nfstream package," 2021. [Online]. Available: <https://www.nfstream.org/>
- [46] Wireshark, "Tshark - a terminal oriented network protocol analyzer," https://www.wireshark.org/docs/wsug_html_chunked/AppToolstshark.html, 2021.
- [47] N. Hason, A. Dvir, and C. Hajaj, *Robust Malicious Domain Detection*, 06 2020, pp. 45–61.
- [48] L. Orevi, A. Herzberg, and H. Zlatokrilov, "Dns-dns: Dns-based de-nat scheme," in *Cryptology and Network Security*, Cham, 2018, pp. 69–88.
- [49] H. Chen and T. Lin, "The challenge of only one flow problem for traffic classification in identity obfuscation environments," *IEEE Access*, vol. 9, pp. 84 110–84 121, 2021.
- [50] C.-H. Hung and H.-M. Sun, "A botnet detection system based on machine-learning using flow-based features," *SECURWARE*, 2018.
- [51] S. Rezaei, B. Kroencke, and X. Liu, "Large-scale mobile app identification using deep learning," *IEEE Access*, vol. 8, pp. 348–362, 2020.
- [52] I. Lee, H. Roh, and W. Lee, "Encrypted malware traffic detection using incremental learning."
- [53] J. Althouse, "Ja3 and ja3s - tls fingerprinting," 2019. [Online]. Available: <https://engineering.salesforce.com/tls-fingerprinting-with-ja3-and-ja3s-24736285967>
- [54] "Tensorflow," Aug. 2021. [Online]. Available: <https://www.tensorflow.org/>
- [55] TensorFlow, "Loading image dataset from directory," 2021. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/utils/image_dataset_from_directory
- [56] Y. Zhang, S. Zhao, J. Zhang, X. Ma, and F. Huang, "STNN: A novel TLS/SSL encrypted traffic classification system based on stereo transform neural network," in *ICPADS, Tianjin, China, December 4-6, 2019*, pp. 907–910.
- [57] I. Akbari, M. A. Salahuddin, L. Ven, N. Limam, R. Boutaba, B. Mathieu, S. Moteau, and S. Tuffin, "A look behind the curtain: Traffic classification in an increasingly encrypted web," *Proc. ACM Meas. Anal. Comput. Syst.*, 2021.
- [58] O. Bader, A. Lichy, C. Hajaj, R. Dubin, and A. Dvir, "Framework evaluation sample code on GitHub." [Online]. Available: <https://github.com/ArielCyber/OSF-EIMTC/examples>
- [59] D. Yadav, R. Jain, H. Agrawal, P. Chattopadhyay, T. Singh, A. Jain, S. B. Singh, S. Lee, and D. Batra, "Evalai: Towards better evaluation systems for ai agents," 2019.
- [60] D. Yadav, R. Jain, H. Agrawal, P. Chattopadhyay, A. J. Taranjeet Singh, S. B. Singh, S. Lee, and D. Batra, "EvalAI: Towards better evaluation systems for ai agents." [Online]. Available: <https://eval.ai/>

APPENDIX A

ARIEL DATASET ANALYSIS

The figures for this dataset are depicted in figure 6. It is apparent that not all browsers share the same set of applications. For example, *Facebook* (fb) is only shared among *Chrome* and *Firefox* labels. Furthermore, *Ubuntu* is the only operating system that hosted that combination of browser and application. A similar observation on the different operating systems (colored in bars) is that some applications and browsers can only be found in the data among specific operating systems, some for technical reasons such as *Internet Explorer* and *Safari*, which are exclusively for *Windows* and *OSX* operating systems respectively. From figure 6a it is evident that across all labels, with the exception of *c-tweet* (*Windows*), the packet size was larger on average in the download flow than the upload flow, This trend can also be seen in figures 6b and 6c illustrating the average number of packets and bytes per session. Note that there are some non-web applications that operated in the background while

recording, which are *Dropbox*, *Microsoft* (services), *Vine*, and *Teamviewer*, hence, they do not contain a browser label.

APPENDIX B ISCX2016 DATASET ANALYSIS

The figures for this dataset are depicted in figure 7. While most of the applications can be found in both *VPN* and *non-VPN* traffic, there is a relatively small number of applications that are exclusive to a one category of encapsulation (only *non-VPN*). Namely *Gmail*, *scp*, *Facebook*, and the video portions of *Hangouts* and *Skype*. In a swift glance at the error lines in figures 7b, 7c, 7d, and 7e, it is evident that the standard deviation is very high for the number of packets, bytes, duration, and inter-arrival time per session, hinting that each feature on its own might not be a most excellent separator between the classes. From the two figures of protocol frequencies per label, for *non-VPN* 7f and for *VPN* 7g, the sessions with protocols other than UDP and TCP are more apparent in *non-VPN* than *VPN* encapsulated traffic. Both in frequency and in the set of different protocols. A researcher that adopt this dataset might want to clean the less important protocols while attempting to classify a network flow to its traffic type and application.

APPENDIX C MAPPGRAPH DATASET PORTION ANALYSIS

The figures for this dataset are depicted in figure 8. The flow-based protocol distribution of TCP and UDP (in figure 8f is almost even across each application, there is no application that is drastically more UDP or TCP based. Which is unique to this dataset portion when compared to the other datasets we have analyzed. In all graphs, except for average packet size (fig 8a), it is prevalent that the standard deviation is high, which might be related to the difference in statistical characteristics between UDP and TCP sessions. From figure 8g we see that *Skype* has the most unopened TCP sessions with 159 sessions, which amounts to 27.6% of the total TCP sessions of *Skype*. While on the other end there is *Soundcloud* with a single unopened TCP session.

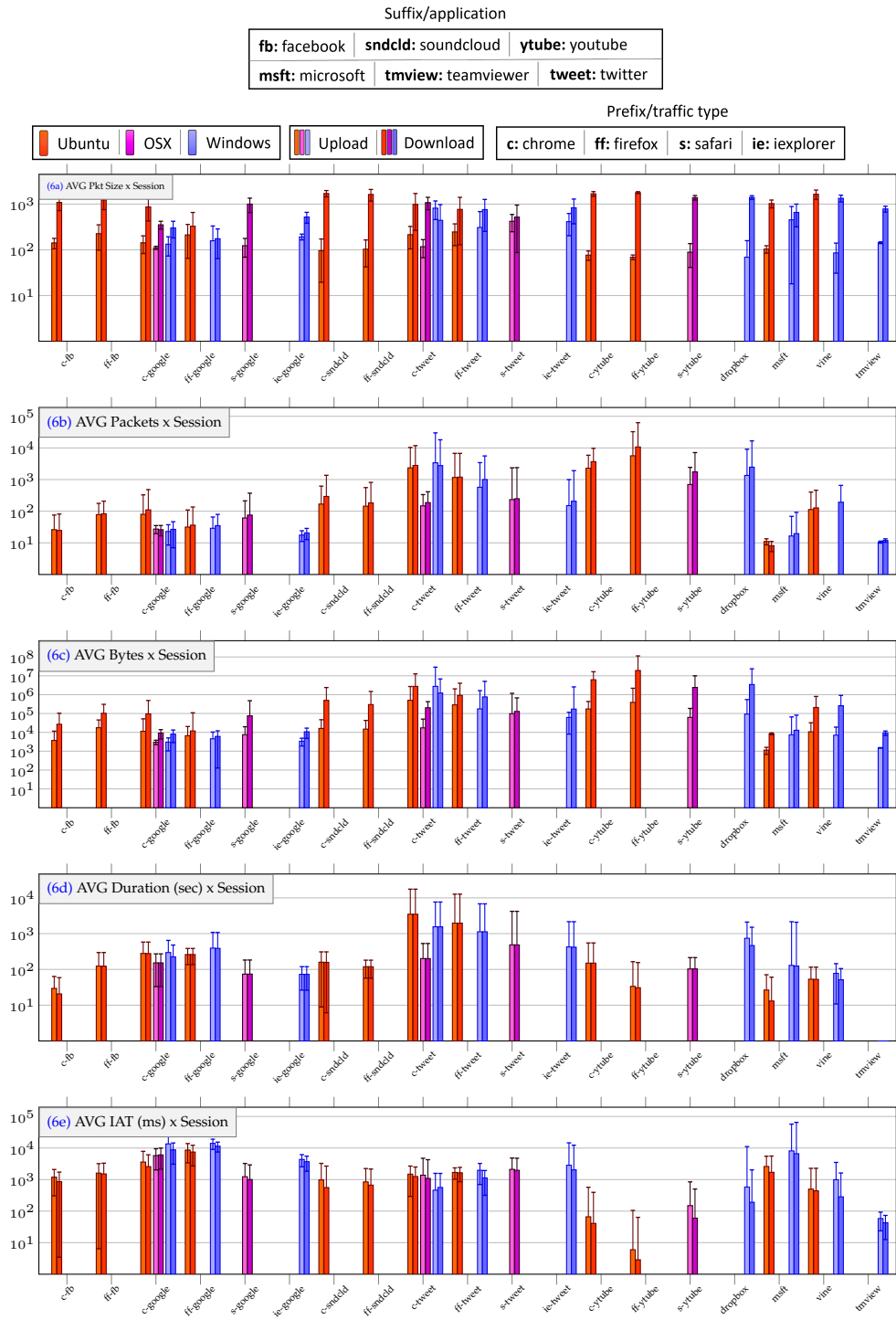


Fig. 6: Analysis of Ariel (BOA2016).



Fig. 7: Analysis of ISCX2016.

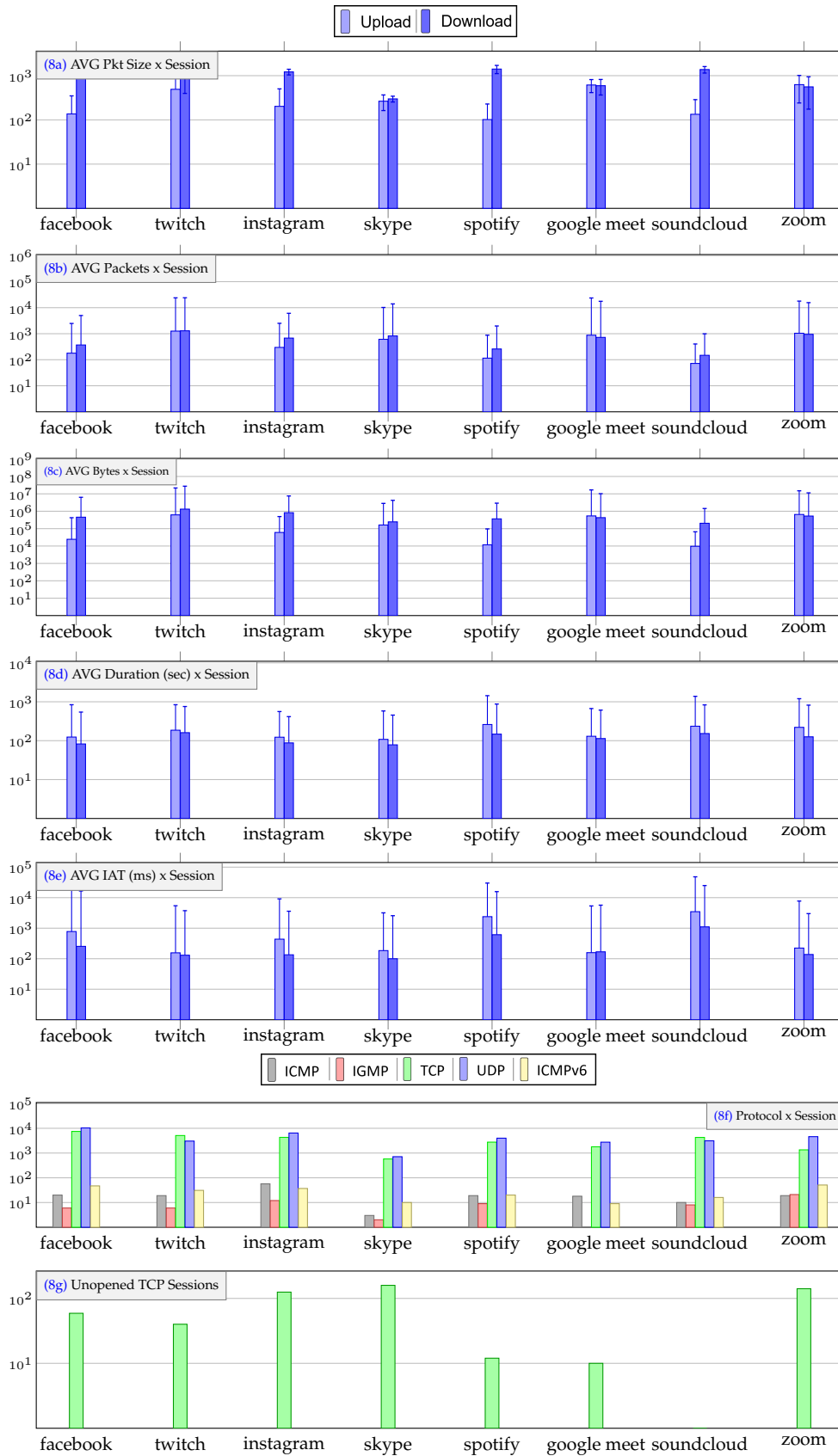


Fig. 8: Analysis of a portion of MAppGraph dataset (of 8 applications).