AI/ML and Cybersecurity: The Emperor has no Clothes
Technical Report

Abstract

Network practitioners remain highly reluctant to relying on Artificial Intelligence (AI) or Machine Learning (ML) to solve real-world problems and deploy AI/ML-based solutions in their production networks. To shed light on this reluctance, we closely examine some recently published cybersecurity-related studies that are fully reproducible. Instead of the touted success stories, we show that these studies indicate a deep-rooted problem with AI/ML and cybersecurity that is succinctly captured by the expression "the emperor has no clothes." At best, the published learning models lack basic scientific rigor in demonstrating that they are credible (i.e., generalize as expected in deployment scenarios). At worse, many of them are broken beyond repair in that they are, for example, the result of simple "shortcut learning." To ensure that the much-heralded AI/ML technologies can achieve their full potential when applied in the cybersecurity domain, we argue for the need to revisit how AI/ML is used in this domain. We propose a new two-pronged approach (and associated new AI/ML pipeline) that emphasizes **explainable** rather than black-box learning models and requires published AI/ML research artifacts to be fully **reproducible**. We demonstrate the benefits of the proposed approach for addressing the rampant problem of underspecification of trained AI/ML models (e.g., cases of overfitting or shortcut learning) and discuss its implications on the future use of AI/ML for cybersecurity in theory and in practice.

I. Introduction

Compared to computer vision or autonomous car technology, where the use of Artificial Intelligence (AI) and Machine Learning (ML) has been adopted early on and with enormous success, the networking area has been relatively late in joining the AI/ML fray. Importantly, network operators and cybersecurity experts have remained doubtful about adopting AI/ML-based solutions and deploying them in their production networks, be it to help with network performance-related issues or address network security-specific problems.

In this paper, our focus is on the application of AI/ML to cybersecurity and on the challenges that this application domain poses. For one, there exists an obvious mismatch between the black-box nature of some of the most commonly considered AI/ML models and what network practitioners expect from or look for in a new technology like AI/ML. That is, while black-box learning models are inherently incapable of providing insights into their "inner workings" or underlying decision-making process, network operators and cybersecurity experts are particularly keen on gaining a basic understanding of how these newly proposed models work in practice so they can be trusted upon deployment in real-world production settings.

At the same time, AI/ML research has, in general, paid little to no attention to reproducibility (i.e., the ability of third parties to independently assess and validate relevant research artifacts), contributing to and accelerating an arguable reproducibility crisis in science [41], [43], [42]. This lack of reproducibility has also played an important role in the general distrust that network practitioners have of using AI/ML-based solutions in practice. Importantly, it prevents the basic evaluation and assessment of whether or not newly-developed AI/ML research artifacts do generalize across different network environments or are only valid for the specific setting from which the underlying training data was obtained in the first place. At the same time, many reproducibility efforts suffer from a basic and widely-lamented lack of relevant data in general and labeled data in particular. This is especially true for an application domain like cybersecurity, where suitable training datasets often contain highly sensitive information that generally prevents third-party sharing of the raw data. While certain anonymized versions of the data may be suitable for sharing, most data anonymization efforts result in a loss of potentially valuable information, which often impedes subsequent reproducibility efforts.

We argue in this paper that for AI/ML to be embraced by network practitioners and have an impact in practice, new research efforts are needed to re-invent how AI/ML ought to be used in an application domain like cybersecurity. In particular, we propose a two-pronged approach that focuses squarely on addressing the trust issue and comprises the following two critical efforts: (i) Developing and evaluating **explainable or interpretable AI/ML models** that can take the form of simple decision trees purposefully extracted from commonly-considered black-box models [3]; and (ii) Requiring the **reproducibility of developed AI/ML research artifacts** so as to ensure safe third-party sharing of newly-designed learning models/algorithms and avoid the problem-ridden third-party sharing of sensitive data altogether [11].

To implement our proposed approach in practice, we develop a new AI/ML pipeline that details the workflow for how modern AI/ML should be used in the application domain of cybersecurity. This newly proposed AI/ML pipeline is our answer to the commonly-expressed need for gaining insight into the operations of newly-proposed black-box learning models that are intended to be used in practice as part of important decision-making processes. The new AI/ML pipeline builds on the standard pipeline currently in use but extends it in significant ways by elevating extracted white-box learning models to the role of first-class citizens. Such a role is not only justified by the ability of explainable learning models to instill in network practitioners the required level of trust but also by the critical capacity that these types of models have for studying the well-documented problem of *underspecification* of developed black-box learning models [9].

For a given black-box learning model, underspecification refers to the problem of determining whether the learning model's excellent performance (e.g., high prediction accuracy) is indeed due to its ability to encode essential structure of the underlying system or simply the result of some inductive bias encoded by the trained model. By applying our newly proposed AI/ML pipeline, we illustrate with examples of published black-box learning models that have been developed for cybersecurity-related problems how extracted white-box learning models become potent tools for revealing the presence of such inductive biases in the underlying black-box model, especially when these biases are the results of unintended learning strategies such as shortcut learning (also known as "Clever Hans" effect [48], [49]), spurious correlations or overfitting [10]. In the process, we also demonstrate how suitably extracted white-box learning models can be used to determine whether the corresponding original black-box learning model is *credible*; that is, generalizes as expected in deployment scenarios and can hence be trusted to not only work in theory (*i.e.,* test set is independent and identically distributed with regard to the training set) but also in practice (*i.e.,* test set is out-of-distribution with regards to the training set).

Finally, while we are not the first to impress upon networking researchers the critical need to ensure that their published AI/ML research artifacts be reproducible by third parties (*e.g.,* see [11] and references therein), we are also not the first to comment on the unique challenges that the cybersecurity area poses as an application domain for AI/ML. In fact, not only do we fully agree with much of previously-expressed skepticism about the use of AI/ML for cybersecurity-related problems (e.g., see [21], [24]), but we argue in this paper more forcefully that when it comes to the application of AI/ML in its current form for cybersecurity, "the emperor has no clothes." More constructively, the purpose of this paper is to "provide the emperor with at least some clothes," ordinary or unimpressive as they may be.

II. Background and Related Work

After briefly summarizing past research efforts on the application of AI/ML for cybersecurity, we describe in this section relevant work on explainable AI/ML and reproducible AI/ML research that motivates of our proposed approach.

A. *On the Use of AI/ML for Cybersecurity*

A number of recent survey articles have provided extensive overviews of the growing research literature on the use of AI/ML in the field of networking; e.g., see [27], [28]. Other survey articles focus specifically on the application of AI/ML for cybersecurity (*e.g.,* see [25], [24], [23], [26], with some specializing even further by only considering publications on the use of AI/ML for network intrusion detection (*e.g.,* [21], [22]). The potential implications of AI/ML for cybersecurity have also been the topic of a recent workshop report by the U.S. National Academies of Science [29]. Many of these articles comment on why despite extensive academic research efforts, AI/ML technology is rarely used in real-world production networks and discuss some of the unique challenges that the area of cybersecurity poses as an application domain for AI/ML [21], [24]. In this paper, we go beyond critiquing the use of AI/ML for cybersecurity and present a constructive approach that provides a much-needed framework for rigorously evaluating, assessing, and comparing the AI/ML artifacts (*e.g.,* learning algorithms, training data, and learning models) produced by the academic and industry researchers.

B. *Explainable AI/ML*

To encourage new research efforts into developing "explainable" models that enable human users to understand, trust, efficiently diagnose and effectively manage the emerging generation of AI/ML-based downstream applications, DARPA launched in 2017 a new program on "Explainable Artificial Intelligence (XAI)" [2]. Subsequent efforts by the AI/ML research community include new methods for extracting explainable or "white-box" models from commonly-considered black-box learning models (e.g., see [34], [38], [35], [3]) and for examining the validity of obtained black-box explanations (e.g., see [40], [36]).

The basic idea of these efforts is to approximate a complex black-box model using a simple and explainable learning model such as a decision tree. Assuming the approximation quality or "fidelity" of such an extracted decision tree is sufficiently good, then the extracted decision tree becomes the main vehicle for studying the decision-making process of the underlying black-box model and examining its properties. In this paper, we propose a change to how AI/ML is currently used for cybersecurity and require that the development of effective black-box models is no longer the primary objective but just a necessary step towards deriving and assessing high-quality explainable learning models that invite interrogation by domain experts.

C. *Reproducible AI/ML Research*

Computer vision researchers have *ImageNet* [13], [14], and researchers working on autonomous driving technology have increasingly access to datasets provided by the various commercial driverless technology companies (*e.g.,* NuScenes [15], ArgoVerse [16], Waymo Open Dataset [17]). Access to these open-source datasets is widely credited with invigorating reproducible AI/ML-related research in these areas. In stark contrast, fledgling reproducibility efforts in the cybersecurity area quickly reduce to demands for data sharing in an application domain of AI/ML where data is hard to come by in the first place. Such data sharing requirements have severely hamstrung reproducible AI/ML research in the cybersecurity domain. To overcome these challenges, we argue in this paper for a paradigm shift in addressing both the data problem and reproducibility crisis afflicting AI/ML-based cybersecurity research. In particular, we assert that academic researchers in the area of AI/ML and cybersecurity (and beyond) accept the lack of open-source datasets and a general reluctance for widespread data sharing as faits accomplis, and instead embrace the sharing of AI/ML research artifacts such as considered learning algorithms or developed learning models, and take charge of the data problem by collecting the necessary data themselves, in close collaboration with their universities' IT organization (see also [11]).

III. What AI/ML for Cybersecurity?

In this section, we motivate our approach to reinventing AI/ML for cybersecurity and present our new AI/ML pipeline.

A. *On the "Old" AI/ML for Cybersecurity*

To get a sense of the lay of the land, we performed a limited survey of the existing literature on applications of AI/ML to cybersecurity-related problems. We grouped the encountered efforts according to their ability to be reproducible and identified three different categories:

Learning model is published and training data is publicly available. Only a very small number of studies reported in the existing literature fall in this category, including [50], [51] to mention a few. We commend these studies because third-party researchers can carefully scrutinize the research artifacts that make up these efforts. At the same time, we encountered hardly any studies that made use of the afforded reproducibility opportunities and performed a careful examination of the published research artifacts (see Section 4).

Learning model is published and training data is not publicly available. This category makes up the vast majority of studies reported in the existing literature. Unfortunately, for competitive, legal, or other reasons, the use of these efforts' training datasets is typically confined to the very researchers who developed the published learning models in the first place, which makes reproducibility of the described research artifacts by third-party researchers impossible (*e.g.,* see [69], [70], [71], [72], [73]).

Learning model is not published and training data is not publicly available. This category consists mostly of commercial solutions. These solutions are marketed by an ever-growing number of cybersecurity companies that claim to leverage AI/ML-based technologies in their products. The proprietary nature of these products rules out any scientifically rigorous evaluation by third parties and prevents even the most basic attempts at reproducibility (e.g., see [55], [56]).

The two main takeaways from this limited survey of the existing literature on the application of AI/ML for cybersecurity are (1) reproducibility efforts are the exception rather than the rule, and (2) the literature on AI/ML for cybersecurity is void of efforts that apply explainable AI/ML, at least at a level comparable to the use case we describe in detail in Section 3.2 below. Importantly, the general lack of reproducibility efforts makes it impossible to carefully scrutinize most of the published learning models developed with cyber security-specific tasks in mind. This observation highlights the current precarious state-of-the-art in AI/ML-based cybersecurity research where the reported (typically superb) performance of learning models across all three categories cannot be taken at face value and where their ability to generalize as expected in deployment scenarios remains at best unknown and is at worse more than questionable.

B. An Illustrative Use Case

In the absence of a compelling published study that concerns an application of modern AI/ML to a cybersecurity problem and allows us to illustrate the issues we raise in this paper, we find inspiration in a recent effort that concerns the use of modern AI/ML for a network performance problem.

The effort in question deals with a recently proposed AI/ML-based system for adaptive bitrate (ABR) video streaming and comprises three different published papers [5], [7], [8]. The original paper [5] presents Pensieve, a new black-box learning model that combines (deep) neural network models with reinforcement learning (RL) to learn an optimized control policy for bitrate adaption in a data-driven and automatic manner. From our perspective, the most relevant aspect of this work is that the authors of [5] open-sourced all Pensieve-related research artifacts, including data, learning model, and code at [6] to encourage full reproducibility of their work. A subsequent paper [7] describes an initial attempt at "cracking open" Pensieve's black-box learning model and relies critically on the ability to reproduce the artifacts developed in [5]. By applying a set of previously developed techniques that are commonly referred to as *local interpretability/explainability* tools (*e.g.,* see [46], [47]), the authors of [7] illustrate several findings that show the potential that local explainability tools have in increasing the network operators' trust in AI/ML-based systems like Pensieve.

Unfortunately, local explainability tools like the ones employed in [7] are limited in their capabilities and fall short in explaining the behavior of a given black-box learning model as a whole (*i.e., global* *explainability*). The main contribution of the third paper [8] is in exploring the use of such global explainability tools to examine Pensieve's black-box learning models. To this end, the authors of [8] build on recent advances in the area of explainable AI/ML (*e.g.,* see [34], [38]) and extract a "white-box" learning model in the form of a decision tree from the black-box learning model used by Pensieve. A basic examination of the resulting highly structured and eminently interpretable decision tree demonstrates its ability to explain Pensieve's black-box learning model as well as the learned policies extracted by this method.

In summary, the Pensieve example [5] convincingly demonstrates the merits of reproducible AI/ML-based networking research. The subsequent efforts, which first used local explainability tools to "look under the hood" of Pensieve's black-box learning model [7] and culminated in applying a global explainability method capable of distilling Pensieve's incomprehensible but high-performing learned policies into simple and highly interpretable decision tree policies [8], could only succeed thanks to the open-source nature of the research artifacts that the authors of [5] released as part of their work. However, we show in the rest of the paper that there is more to explainability than these Pensieve-specific efforts indicate.

C. A New AI/ML Pipeline

Motivated and inspired by the described body of Pensieve-related efforts [5], [7], [8], we next propose a new workflow for how modern AI/ML be used in the application domain of cybersecurity. We describe the proposed workflow in terms of a new AI/ML pipeline that elevates suitably extracted white-box learning models to the role of first-class citizens.

- **Step 1**: Defining a specific training task and its description in terms of a chosen model specification or algorithm;
- **Step 2**: Using a provided training dataset and the selected algorithm to train a black-box learning model;
- **Step 3**: Evaluating the obtained learning model by traditional procedures that measure the model's performance using a dataset (*i.e.,* typically using a train-test split) that has the same distribution as the given training dataset;
- **Step 4**: Extracting a high-fidelity, post-hoc white-box learning model (*e.g.,* decision tree) from the black-box learning model obtained in Step 2 and evaluated in Step 3; and
- **Step 5**: Examining the white-box learning model obtained in Step 4 to determine whether or not the black-box learning model obtained in Step 2 and evaluated in Step 3 is credible and can be trusted.

In effect, the first three steps of this new AI/ML pipeline comprise the standard AI/ML pipeline currently in use, and it is the newly added Steps 4-5 that capture the requirement of a growing number of AI/ML applications to real-world problems (*e.g.,* cybersecurity). This requirement demands to move beyond the single-issue objective of creating high-performing black-box learning models and strive instead for a more holistic assessment of them that focuses on aspects commonly associated with "trust", including interpretability, credibility, and effectiveness. To this end, Step 4 is intended to produce an appropriate post-hoc white-box learning model that, for all practical purposes, serves as a high-quality substitute for the underlying high-performing but incomprehensible black-box learning model. This explainable learning model is then used in Step 5 to examine important properties (*e.g.,* fairness, safety, robustness) of the obtained black-box learning model or its learned policies. In particular, we view this modified AI/ML pipeline as a practical means to provide transparency of and establish trust in high-performing black-box models to the point where it can become a main vehicle for answering questions concerning the ethical quality of black-box based decision making [45].

IV. The Modified AI/ML Pipeline "in Action": Preliminary Results

This section illustrates with two concrete example how a previously published study which is reproducible and concerns a specific cybersecurity problem fares when scrutinized with the help of the newly-proposed AI/ML pipeline.

A. A Case of Shortcut Learning

Selected publication: We consider the paper [50] that presents an AI/ML-based approach for encrypted traffic classification. The proposed approach integrates feature design, feature extraction, and feature selection into a common framework and uses one-dimensional convolutional neural network (1D-CNN) models, prime examples of Deep Neural Networks (DNN), to automatically learn the relationship between the raw input data and the expected output labels. For the particular problem of classifying VPN vs non-VPN traffic, the authors use the public PCAPs of the ISCX VPN-nonVPN dataset [64] to train a 1D-CNN learning model, treating the packet-level traffic associated with each session as a 2D image of size 28x28. As a result, the proposed model views input traffic samples as discrete byte streams of fixed length (i.e., 784 bytes) and treats each byte as a possible "feature". The authors report outstanding performance (i.e., 100% (99.9%) precision and 99.9% (100%) recall for Non-VPN (VPN) traffic). All AI/ML research artifacts and datasets [53] are available at [54], allowing full reproducibility of the described models and reported findings.

Key question: Is the reported learning model credible? To examine this problem, we first apply the new AI/ML pipeline presented in Section III-C. In particular, by first performing Steps 1-3, we reproduced the black-box learning model (i.e., 1D-CNN) and the results presented in [50, Table VI] for the problem of classifying VPN vs non-VPN traffic. Next, in Step 4, we extracted a decision tree from the black-box 1D-CNN learning model using a method derived from [35]. We show the resulting decision tree in Figure 1 and explain it in detail below. We illustrate with several experiments the type of efforts that are required to examine the credibility of a given black-box model by means of investigating a suitably extracted decision tree model (i.e., applying the final Step 5 of the new AI/ML pipeline).

First, to assess the fidelity of the extracted decision tree (*i.e.,* determining how well this extracted white-box model reproduces the decisions made by the black-box model), we used it to classify the test cases from [50], and compared the results with the classification from the black-box, measuring precision, recall, and F-1. To our surprise, despite its small size, the extracted decision tree reproduced all black-box decisions, achieving a perfect F-1 score.

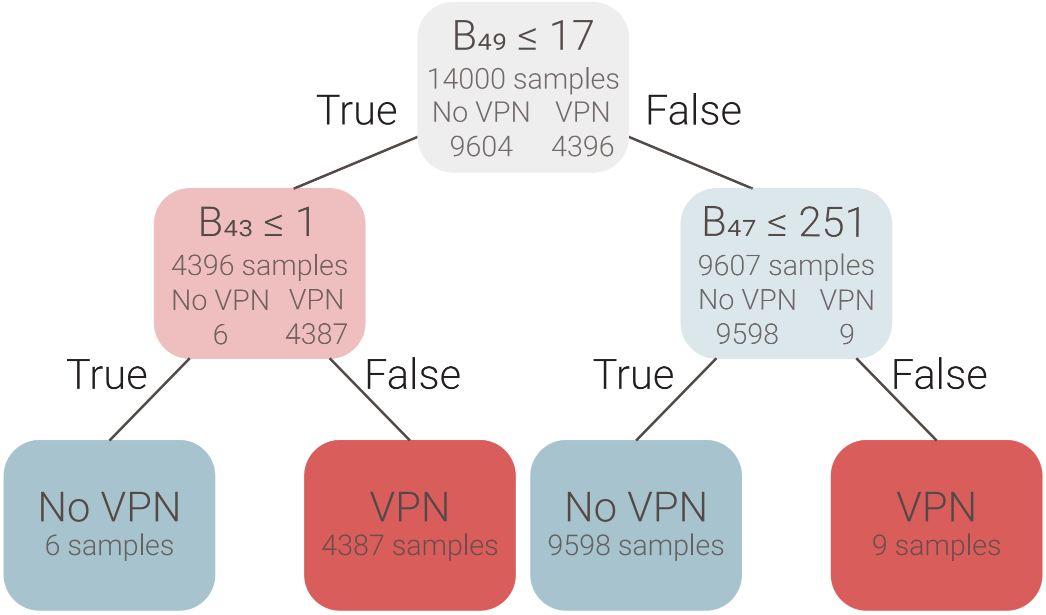
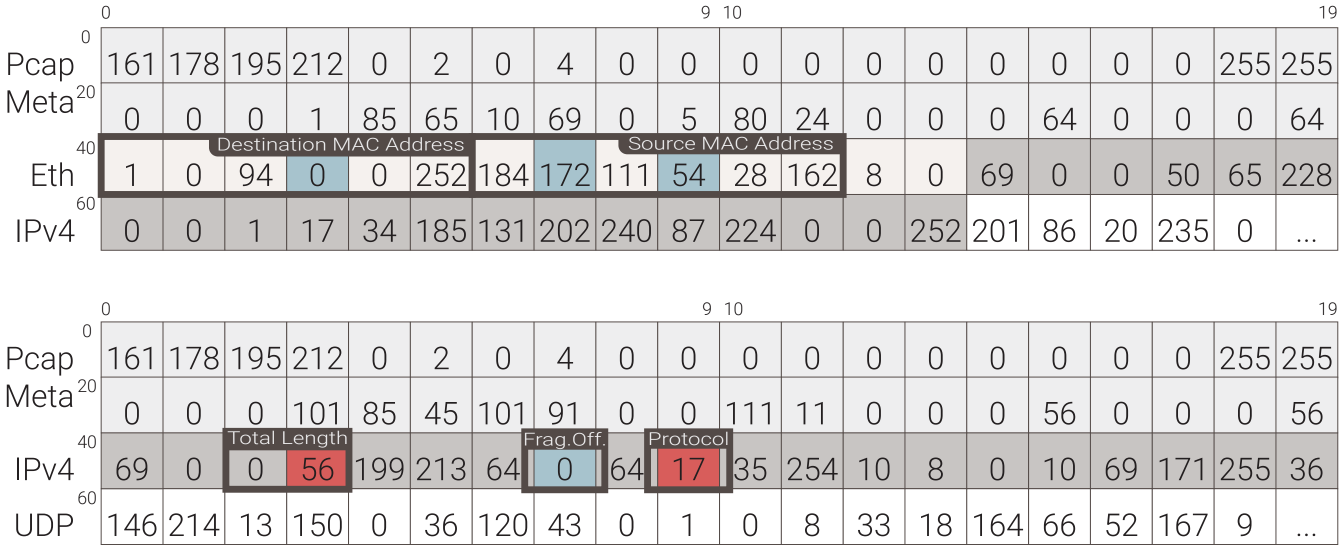


Fig. 1: Decision tree explanation extracted from the blackbox 1D-CNN model, with fidelity (i.e., F-1 score) = 1. Each node depicts a decision based on the indicated byte or feature used as input for the black-box model.

Given this high-fidelity decision tree model, we next analyze its decisions. Correctly interpreting this extracted decision tree requires knowing in detail how the traffic data is organized before it is ingested by the black-box learning model. Since the decision tree‘s classification is based on only three bytes or input features in the initial segment of each input sample (i.e., bytes B_{49}, B_{43}, and B_{47}), we inspected samples of VPN and Non-VPN test cases to uncover the “meaning” of those bytes and show in Figure 2 a schematic view of the first 80 bytes of actual input data used in [50]. We notice that each input sample consists of an initial set of bytes that contain PCAP metadata, Ethernet header information, and IP header data. Noticeably, none of these initial bytes or “features” say anything about actual VPN or non-VPN traffic.

Upon further scrutiny of the public dataset [53], we noticed that while non-VPN traffic samples always contain Ethernet headers, roughly 90% of the VPN traffic samples have no Ethernet header. As shown in Figure 2, for each non-VPN input sample, the first 40 bytes consist of PCAP metadata, followed by 14 bytes of the Ethernet header and 20 bytes of the IP header, with the remaining bytes representing TCP or UDP headers and (encrypted or padded) application data. In contrast, each VPN input sample starts with 40 bytes of PCAP metadata, followed by 20 bytes of the IP header, then TCP or UDP headers, and application data. Thus, if B_k denotes the byte in position k (i.e., k-th feature), then for $k > 20$, there is clearly a misalignment of features between different input samples, resulting in the same byte k having completely different semantics in the two types of traffic.

With this understanding of the input data, the extracted decision tree is easy to explain and is consistent with available domain knowledge. From Figure 1, we see that feature B_{49} was used as the splitting



		0									9	10									19
Pcap	0	161	178	195	212	0	2	0	4	0	0	0	0	0	0	0	0	0	0	255	255
Meta	20	0	0	0	1	85	65	10	69	0	5	80	24	0	0	0	64	0	0	0	64
Eth	40	1	0	94	0	0	252	184	172	111	54	28	162	8	0	69	0	0	50	65	228
IPv4	60	0	0	1	17	34	185	131	202	240	87	224	0	0	252	201	86	20	235	0	...

		0									9	10									19
Pcap	0	161	178	195	212	0	2	0	4	0	0	0	0	0	0	0	0	0	0	255	255
Meta	20	0	0	0	101	85	45	101	91	0	0	111	11	0	0	0	56	0	0	0	56
IPv4	40	69	0	0	56	199	213	64	0	64	17	35	254	10	8	0	10	69	171	255	36
UDP	60	146	214	13	150	0	36	120	43	0	1	0	8	33	18	164	66	52	167	9	...

Fig. 2: First 80 bytes from the training dataset for non VPN (top) and VPN (bottom). Cells in red and blue contribute to a VPN and No-VPN classification, respectively, according to the Decision Tree in Figure 1.

criterion in the root node of the decision tree. Due to the mentioned feature misalignment between VPN and Non-VPN test cases shown in Figure 2, the semantic value of B_{49} in VPN samples is the protocol byte in the IPv4 header, while for Non-VPN samples, the feature is the fourth byte of the source MAC address of the Ethernet header. Since the VPN traffic in the dataset only uses UDP (B_{49} = 17) or TCP (B_{49} = 6), the first root node of the decision tree splits almost the entire sample set by comparing the protocol byte used in VPN traffic with the fourth byte of random MAC addresses of the machines used to generate the traffic trace. Coincidentally, we found only two machines used in Non-VPN traces which had the MAC address (`54:9f:35:0d:e9:c2` and `2c:44:fd:02:16:ef`) with the value of the fourth byte less than or equal to 17, making feature B_{49} a prime "shortcut" to classify the traffic.

Following the left branch of the decision tree that has most VPN samples, we see that to weed out the few remaining samples of Non-VPN traffic, byte or feature B_{43} was used as the splitting criterion. Again, because of the feature misalignment, in most VPN samples, feature B_{43} corresponds to the total length of the IP packet, while for Non-VPN samples, this feature represents the fourth byte of the destination MAC address in the Ethernet header. The splitting criterion shows that the decision tree classifies the traffic by comparing the total length of the VPN IP packets (never less than one for the entire trace) with the fourth byte of the destination MAC addresses of the machines that generated the Non-VPN dataset samples, which were always 0. Once again, the black-box model is taking a shortcut to classify VPN vs Non-VPN traffic.

Following the right branch of the decision tree that has most Non-VPN samples, we notice that feature B_{47} is used to distinguish the remaining VPN samples. By scrutinizing the few VPN traffic samples in

TABLE I: Average precision, recall and F-1 score of black-box classifier of VPN vs Non-VPN traffic.

Validation dataset	Avg. Precision	Avg. Recall	Avg. F1
Untampered	0.959	0.956	0.955
Tampered-43-47-49	0.959	0.956	0.955
Tampered-32-to-63	0.889	0.861	0.856
Tampered-0-to-63	0.831	0.757	0.734
Tampered-0-to-127	0.753	0.555	0.398

this branch, we noticed that they fall into the 10% of VPN sessions captured **with** the Ethernet headers, making this decision a direct comparison between the second bytes of source MAC addresses. However, once again, by sheer coincidence, the destination MAC addresses of these samples had the second byte set to 255 (*e.g.,* `00:ff:c2:e4:8c:db`).

Even though the extracted decision tree is a high-fidelity proxy for the 1D-CNN black-box model, it is unreasonable to expect that a simple 3-node structure encompasses the model's entire decision-making process. We verify this intuition by generating a tampered validation dataset for the black-box model. In particular, we changed bytes 43, 47, and 49 in the VPN samples to mimic random Non-VPN samples. By following the logic of the decision tree branches, the black-box model would misclassify all VPN samples. The first two rows of Table I show the average precision, recall, and F-1 score for both classes (VPN vs Non-VPN) for original and tampered datasets. The results show that tampering with only these three features out of 748 had no significant impact on the classification accuracy of the black-box model. However, by performing detective work similar to the one described above, we observed that the black-box model succeeds in finding alternative "shortcut" that are as easy to identify and explain as the previously observed shortcuts.

To further demonstrate that the black-box learning model described in [50] and claimed to be highly successful in learning to classify encrypted VPN and non-VPN traffic is in fact not a credible predictor, we performed additional experiments by tampering with entire ranges of bytes instead of individual bytes. As Table I shows, tampering with byte ranges of 32-64, 0-64, and 0-128 makes it increasingly more difficult for the black-box model to identify alternative shortcut predictors, and not surprisingly, the model's performance (i.e., accuracy) gets worse and quickly reaches the point where, without being able to resort to shortcut learning (*i.e.,* randomly altering the first 128 bytes, which is less than 18% of the features), the model's performance becomes comparable to that of flipping a fair coin.

Answer: The black-box model considered in [50] is not credible as it is a textbook example of "shortcut learning" [10].

B. A Case of IID vs OOD

Selected publications: We consider [51], which presents the publicly available dataset CIC-IDS-2017 with labeled attack traces, and the publications that rely on this dataset to propose AI/ML-based intrusion detection systems. The CIC-IDS-2017 dataset contains traces of benign background traffic and 13 different attacks, such as Heartbleed, DDoS, and PortScans. The dataset also includes a set of 78 pre-computed flow features, such as flow duration, mean Inter Arrival Time (IAT), and mean packet length. Several research efforts report excellent classification results (e.g., average precision and recall above 99% for all classes) of learning models trained on the pre-computed features of the dataset [63], [62], [65], [66], [67].

Key question: Is the reported learning model credible? To answer this question, we again start by applying Steps 1-3 of our AI/ML pipeline to reproduce the reported classification results. To that end, we used the pre-computed features from the dataset to train a Random Forest Classifier, using a 75%-25% train-test split of the data, and were able to reproduce the excellent results reported by several publications. Next, to perform Step 4 of our AI/ML pipeline, we used the same method from the previous section to extract a high-fidelity decision tree from the trained Random Forest Classifier. Finally, in the process of applying the final Step 5, the class of Heartbleed attacks caught our attention and is the focus of this use case. In Heartbleed, an attacker sends a heartbeat message with a value in the size field bigger than the message. A vulnerable server responds with a message with the size equal to the value specified in the size field and reviews information stored locally in its memory [68].

We analyzed the most prominent features used to classify Heartbleed attacks in the decision tree explanation and examined the feature values in the dataset. In doing so, we noticed that in Heartbleed flows, the average backward packet length (*i.e.,* the mean length of response packets from the attacked server) and the backward total IAT features had unusually large values when compared to other classes. Upon further inspection of the attack packet capture files, we realized that the Heartbleed attacks in the generated traffic had a behavior that may not reflect the behavior of a real-world attack – the TCP connection of the attacking flow was never closed. Thus, for each generated Heartbleed flow, a single connection was established with the vulnerable server, and multiple heartbeat messages were sent in that flow to collect compromised data, resulting in an abnormally high backward total IAT and packet length.

Realizing that the dataset contained just one very clear Heartbleed attack pattern, it is not surprising that classifiers trained on this dataset have high accuracy when tested with independent and identically (IID) generated attack samples. However, to demonstrate that such models are credible and generalize as expected in deployment scenarios, it is necessary to also examine them with alternate but realistic test cases, often referred to as out-of-distribution (OOD) samples. To illustrate, we generated 1000 new test cases of Heartbleed attacks, using the same tool described in [51]. However, instead of maintaining an open connection and sending several Heartbeat messages to the vulnerable server, we immediately closed each connection after one Heartbeat request triggered a response with compromised data. As expected, this experiment resulted in flows with much smaller backward total IAT, but with very similar backward packet length (we used similar packet sizes as the original trace).

Finally, we examined the Random Forest Classifier we had trained in steps 1-3 of the AI/ML pipeline when using the newly generated Heartbleed flows as test data. The results were telling: with just a simple change in the attack pattern, the classifier was unable to correctly identify a single one of the 1000 new Heartbleed attacks, resulting in precision and recall of 0. In short, this experiment demonstrates that the considered black-box learning model overfits on the IID cases, is not a credible predictor of realistic

OOD cases, and does not learn anything pertinent to real-world Heartbleed attack.

Answer: No, because the model suffers from the problem of underspecification and is a textbook example of “overfitting”.

V. Discussion

The widely-lamented dearth of data in general and labeled data in particular is widely viewed to be a major reason for why despite the age of “big data” and despite extensive academic research efforts, the use of AI/ML for cybersecurity has produced underwhelming results, and the application of modern AI/ML technology in real-world production networks remains the exception rather than the rule. In this paper, we identify the way AI/ML in its current form is applied to solve cybersecurity-related problems as another major reason for why AI/ML has not been able to “deliver” for cybersecurity. In particular, we argue that for the much-heralded AI/ML technologies to achieve their full potential when applied to cybersecurity, AI/ML research has to focus on the development of white-box rather than black-box models and has to ensure that all relevant AI/ML research artifacts are fully reproducible. To accomplish these two objectives, researchers have to accept the lack of open-source datasets and a general reluctance for widespread data sharing as a given, are asked to instead embrace appropriate sharing of AI/ML research artifacts such as considered learning algorithms or developed learning models, and need to take charge of the data problem by collecting the necessary data themselves.

To realize this “new” AI/ML, we propose a new AI/ML workflow in the form of an extension of the traditional AI/ML pipeline. This proposed “new” AI/ML for cybersecurity has ramifications beyond AI/ML research. For the fast-growing number of commercial cybersecurity companies, the absence of any scientific standards or benchmarks by which the myriad of existing vendor products can be evaluated or compared means that these vendors have little to no incentives to let third parties interrogate their products’ embedded AI/ML artifacts and examine if the products are technically sound and not prone rampant underspecification and its debilitating effects in deployment scenarios. Our proposed new AI/ML pipeline provides a constructive approach and much-needed framework for rigorously evaluating and comparing the AI/ML artifacts produced by academia and industry alike.

However, much work remains before it will be possible to rigorously assess whether or not published learning models or commercial AI/ML-based solutions are credible (i.e., generalize as expected in deployment scenarios) and can be trusted by network practitioners (i.e., the provided explanations for their underlying decision making process are consistent with the experts’ domain knowledge). In particular, identifying concrete use cases for demonstrating that this “new” AI/ML is, in fact, a two-way street and enables both AI/ML models to learn from domain experts (this paper) and domain experts to learn from a black-box model looms as a promising and exciting future research. In effect, the objective is to demonstrate in the context of specific cybersecurity-related problems that a properly trained future learning models routinely make decisions that only become part of an expert’s domain knowledge “after the fact”; that is, after the expert had time to contemplate such model decisions, concluded that they were logical, but had to acknowledge that they were not part of his prior domain knowledge (i.e., a case of an expert learning from a black-box model by means of the proposed new AI/ML pipeline).

VI. Conclusion

In this paper, we criticize the current state-of-the-art in AI/ML in AI/ML-based cybersecurity research and identify a number of major deterrents to applying AI/ML-based cybersecurity in practice. At the same time, we are also constructive by proposing and illustrating a concrete new framework for rigorously evaluating, assessing, and comparing the AI/ML artifacts produced by the academic and industry researchers. In particular, we propose a new AI/ML workflow in the form of a new and purposefully-designed AI/ML pipeline that elevates high-fidelity white-box learning models that have been extracted from an underlying black-box learning model to the role of first-class citizen. As such, the main objective is no longer the development of a high-performance black-box learning model but a careful examination of whether or not a learned model's typically excellent performance (e.g., high prediction accuracy) is indeed due to the model's ability to encode essential structure of the input data or simply the result of some inductive bias encoded by the trained model. It is only by addressing this so-called problem of underspecification that we can advance AI/ML-based cybersecurity research to the point where the credibility of proposed black-box learning models can be rigorously established and where their "inner workings" can be explained to network practitioners.

References

[1] Supplemental material. Available at https://github.com/anon4papers/emperor.

[2] M. Turek. Explainable Artificial Intelligence (XAI). https://www.darpa.mil/program/explainable-artificial-intelligence

[3] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). In: IEEE Access (2018). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466590

[4] https://trustml.github.io/docs/aies20.pdf

[5] H. Mao, R. Netravali, and M. Alizadeh. Neural adaptive video streaming with Pensieve. In: Proc. ACM SIGCOMM (2017). https://people.csail.mit.edu/hongzi/content/publications/Pensieve-Sigcomm17.pdf

[6] Github repository at https://github.com/hongzimao/pensieve.

[7] A. Dethise, M. Canini, and S. Kandual. Cracking open the black box: What observations can tell us about reinforcement learning agents. In: Proc. ACM SIGCOMM NetAI Workshop (2019). https://sands.kaust.edu.sa/papers/cotbb.netai19.pdf

[8] Z. Meng, M. Wang, J. Bai, M. Xu, H. Mao and H. Hu. Interpreting deep learning-based networking systems. In: Proc. ACM SIGCOMM (2020). https://dl.acm.org/doi/pdf/10.1145/3387514.3405859

[9] A. D'Amour et al. Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395 (2020).

[10] R.Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. arXiv preprint arXiv:2004.07780 (2020). https://arxiv.org/pdf/2004.07780.pdfhttps://www.nature.com/articles/s42256-020-00257-z#citeas

[11] A. Gupta, C.Mac-Stoker, and W. Willinger. An Effort to Democratize Networking Research in the Era of AI/ML. In: Proc. ACM SIGCOMM HotNets (2019). https://dl.acm.org/doi/pdf/10.1145/3365609.3365857

[12] C. Song and A. Raghunathan. Information Leakage in Embedding Models. In: Proc. ACM SIGSAC CCS (2020). https://dl.acm.org/doi/pdf/10.1145/3372297.3417270

[13] ImageNet. http://www.image-net.org

[14] ImageNet Publications. http://image-net.org/about-publication

[15] NuScenes. https://www.nuscenes.org

[16] ArgoVerse. https://www.argoverse.org

[17] Waymo. https://waymo.com/open/

[18] DARPA 1999 Dataset. https://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset

[19] KDD'99 Dataset.http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[20] CIC-IDS-2017 Dataset. https://www.unb.ca/cic/datasets/ids-2017.html

[21] R. Sommer and V. Paxson. Outside the Closed World: On Using Machine Learning For Network Intrusion Detection. In: IEEE Symposium on Security and Privacy (2010). https://personal.utdallas.edu/~muratk/courses/dmsec_files/oakland10-ml.pdf

[22] A. L. Buczak and E. Guven. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. In: IEEE Communications Surveys and Tutorials (2016). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7307098

[23] D. S. Berman, A.L. Buczak, J.S. Chavis, and C.L. Corbett. A survey of deep learning methods for cybersecurity. In: Information (2019). https://www.mdpi.com/2078-2489/10/4/122

[24] G. Apruzzese, M. Coljanni, L. Ferretti, A. Guido, and M. Marchetti. On the Effectiveness of Machine and Deep Learning for Cyber Security. In: Proc. 10th International Conference on Cyber Conflict (CyCon X) (2018). https://ccdcoe.org/uploads/2018/10/Art-19-On-the-Effectiveness-of-Machine-and-Deep-Learning-for-Cyber-Security.pdf

[25] Y. Xin, L. Kong, A. Liu, Y. Chen, Y. Li, H. Zhu, and M. Sao. Machine Learning and Deep Learning Methods for Cybersecurity. In: IEEE Access (2018). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8359287

[26] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed and M. Xu. A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. In: IEEE Access (2020). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9277523

[27] R. Boutaba, M. A. Salahuddin, N. Limam, S. Ayoubi N. Shahriar, F.Strada-Solano, and O.M. Caicedo. A comprehensive survey on machine learning for networking: Evolution, applications and research. In: Journal of Internet Services and Applications, Soringer-Verlag (2018). https://jisajournal.springeropen.com/articles/10.1186/s13174-018-0087-2#Tab24

[28] M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang. Machine Learning for Networking: Workflow, Advances and Opportunities. In: IEEE Network (2017). https://arxiv.org/pdf/1709.08339.pdf

[29] National Academies of Sciences, Engineering, and Medicine. Implications of Artificial Intelligence for Cybersecurity. In: Proceedings of a Workshop. Washington, DC (2019). The National Academies Press.

[30] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In: ITCS (2012).

[31] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FAT (2018). http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[32] S. Burton, L. Gauerhof, and C. Heinzemann. Making the case for safety of machine learning in highly automated driving. In: SAFECOMP (2017). https://link.springer.com/chapter/10.1007/978-3-319-66284-8_1

[33] X. Ma, K. Driggs-Campbell ,and M. J. Kochenderfer. Improved Robustness and Safety for Autonomous Vehicle Control with Adversarial Reinforcement Learning. In: 2018 IEEE Intelligent Vehicles Symposium (2018). https://ieeexplore.ieee.org/document/8500450

[34] O. Bastani, C. Kim, and H. Bastani. Interpretability via model extraction. In: FAT/ML Workshop (2017). https://obastani.github.io/docs/fatml17.pdf

[35] O. Bastani, Y. Pu, and A. Solar-Lezama. Verifiable reinforcement learning via policy extraction. In: Proc. NeurIPS (2018). https://dl.acm.org/doi/10.5555/3327144.3327175

[36] H. Lakkaraju and O. Bastani. "How do i fool you?": Manipulating user trust via misleading black box explanations. In: Proc. AAAI.ACM AIES'20 (2020). https://www.aies-conference.com/2020/wp-content/papers/182.pdf

[37] Z. C. Lipton. The mythos of model interpretability In: ACM Queue (2020). https://queue.acm.org/detail.cfm/?id=3241340

[38] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Interpretable and explorable approximations of black box models. In: ACM SIGKDD (2017). https://arxiv.org/abs/1707.01154

[39] H. Lakkaraju, S.H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In: ACM SIGKDD (2016). https://www-cs-faculty.stanford.edu/people/jure/pubs/interpretable-kdd16.pdf

[40] A. Ghorbani, A.Abid, and J. Zou. Interpretationof neural networks is fragile. In: Proc. AAAI (2019) https://ojs.aaai.org/index.php/AAAI/article/view/4252

[41] P. Gosh. AAAS: Machine learning causing science crisis https://www.bbc.com/news/science-environment-47267081

[42] V. Bajpai et al. The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research In: ACM SIGCOMM CCR (2019). https://dl.acm.org/doi/10.1145/3314212.3314217

[43] Artifact Review and Badging Version 1.1 (2020) https://www.acm.org/publications/policies/artifact-review-and-badging-current

[44] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In: Nature Machine Intelligence (2019). https://www.nature.com/articles/s42256-019-0048-x

[45] L. Piano. Ethical principles in machine learning and artificial intelligence: Cases from the field an possible ways forward. In: Humanit. Soc. Sci. Commun. (2020) https://www.nature.com/articles/s41599-020-0501-9#citeas

[46] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In: Proc. ACM SIGKDD (2016). http://www.csc.villanova.edu/~beck/csc8570/papers/ribeiro.pdf

[47] C. Molnar et al. Pitfalls to Avoid when Interpreting Machine Learning Models. arXiv preprint arXiv:2007.04131 (2020). https://arxiv.org/pdf/2007.04131.pdf

[48] S. Lapuschkin, S. Wäldchen, A. Binder, et al. Unmasking Clever Hans predictors and assessing what machines really learn. In: Nature Communications (2019). https://doi.org/10.1038/s41467-019-08987-4

[49] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller. The Clever Hans Effect in Anomaly Detection. In: CoRR (2020). https://arxiv.org/abs/2006.10609

[50] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In: Proc. IEEE Int. Conference on Intelligence and Security Informatics (ISI) (2017). https://ieeexplore.ieee.org/document/8004872

[51] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani. Towards Generating a New Intrusion Detection Dataset and Intrusion Detection Traffic Characterization. In: Proc. Int. Conference on Information Systems Security and Privacy (ICISSP) (2018).

[52] G. D. Gil, A. H. Lashkari, M. Mamun, and A. A. Ghorbani. Characterization of Encrypted and VPN Traffic Using Time-Related Features. In: Proc. Int. Conference on Information Systems Security and Privacy(ICISSP) (2016).

[53] VPN-nonVPN dataset (ISCXVPN2016). https://www.unb.ca/cic/datasets/vpn.html

[54] Github repository https://github.com/echowei/DeepTraffic.

[55] Darktrace. Darktrace Cyber AI Analyst: Autonomous Investigations. White Paper (2021). https://www.darktrace.com/en/resources/wp-cyber-ai-analyst.pdf

[56] R. Molony. Memorizing Behavior: Experiments with Overfit Machine Learning Models. Crowdstrike Blog (2020). https://www.crowdstrike.com/blog/how-we-trained-overfit-models-to-identify-malicious-activity/

[57] R. Lippmann, R. K. Cunningham, D. J. Fried, I. Graf, K. R. Kendall, S. E. Webster, and M. A. Zissman. Results of the 1998 DARPA Offline Intrusion Detection Evaluation. In: Proc. Recent Advances in Intrusion Detection (1999).

[58] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das. The 1999 DARPA Off-line Intrusion Detection Evaluation. In: Computer Networks (2000).

[59] KDD Cup Data (1999). http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[60] J. McHugh. Testing Intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratories. In: ACM Transactions on Information and System Security (2000).

[61] M. V. Mahoney and P. K. Chan. An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection.In: Proc. Recent Advances in Intrusion Detection (2003).

[62] C. Busse-Grawitz, R. Meier, A. Dietmüller, T. Bühler, and L. Vanbever. pForest: In-Network Inference withRandom Forests.CoRR, abs/1909.05680, 2019.

[63] R. Doriguzzi-Corin, S. Millar, S. Scott-Hayward, J. M. del Rincón, and D. Siracusa. Lucid: A Practical,Lightweight Deep Learning Solution for DDoS Attack Detection. IEEE Transactions on Network and Service Management, 17(2):876–889, 2020.

[64] 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 Proceedings of the 2nd International Conference on Information Systems Security and Privacy - ICISSP, pages 407–414. INSTICC, SciTePress, 2016.

[65] S. Dwivedi, M. Vardhan, and S. Tripathi. An effect of chaos grasshopper optimization algorithm for protection of network infrastructure. Computer Networks, 176:107251, 2020.

[66] I. Sharafaldin , A. H. Lashkari , and A. A. Ghorbani. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy - ICISSP, pages 108–116. INSTICC, SciTePress, 2018.

[67] H. Zhang, L. Huang, C. Q. Wu, and Z. Li. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. Computer Networks, 177:107315, 2020.

[68] Z. Durumeric, F. Li, J. Kasten, J. Amann, J. Beekman,M. Payer, N. Weaver, D. Adrian, V. Paxson, M. Bailey,and J. A. Halderman. The Matter of Heartbleed. In Proceedings of the 2014 Conference on InternetMeasurement Conference, IMC '14, page 475–488,New York, NY, USA, 2014. Association for Computing Machinery

[69] A. Finamore, M. Mellia, M. Meo, D. Rossi. Kiss: Stochastic packet inspection classifier for UDP traffic. IEEE/ACM Trans Netw. 2010; 18(5):1505–15.

[70] N. Jing, M. Yang, S. Cheng, Q. Dong, H. Xiong. An efficient SVM-based method for multi-class network traffic classification. In: Performance Computing and Communications Conference (IPCCC), 2011 IEEE 30th International. IEEE: 2011. p. 1–8.

[71] D. Schatzmann, W. Mühlbauer, T. Spyropoulos, X. Dimitropoulos. Digging into https Flow-based classification of webmail traffic. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement: 2010. p. 322–27.

[72] J. Zhang, C. Chen, Y. Xiang, W. Zhou, Y. Xiang. Internet traffic classification by aggregating correlated naive bayes predictions. IEEE Trans Inf Forensic Secur. 2013; 8(1):5–15.

[73] Y. Sun, X. Yin, J. Jiang, V. Sekar, F. Lin, N. Wang, T. Liu, B. Sinopoli. Cs2p: Improving video bitrate selection and adaptation with data-driven throughput prediction. In: Proceedings of the 2016 conference on ACM SIGCOMM 2016 Conference. ACM: 2016. p. 272–85.