Intelligent Malware Defenses

Azqa Nadeem[1](✉), Vera Rimmer[2], Wouter Joosen[2], and Sicco Verwer[1]

[1] Delft University of Technology, Delft, The Netherlands
{azqa.nadeem,s.e.verwer}@tudelft.nl
[2] imec-DistriNet, KU Leuven, Leuven, Belgium
{vera.rimmer,wouter.joosen}@cs.kuleuven.be

Abstract. With rapidly evolving threat landscape surrounding malware, intelligent defenses based on machine learning are paramount. In this chapter, we review the literature proposed in the past decade and identify the state-of-the-art in various related research directions—malware detection, malware analysis, adversarial malware, and malware author attribution. We discuss challenges that emerge when machine learning is applied to malware. We also identify the key issues that need to be addressed by the research community in order to further deepen and systematize research in the malware domain.

1 Introduction

Over the past two decades, malicious software (malware) has emerged as one of the biggest security threats. AV-test, a security research institute, has reported detecting more than 1000 Million malware samples in 2019[1]. According to Accenture, a malware attack on a company can cost $2.4M on average and can take 50 days to resolve[2]. Anti-Viruses (AVs) are considered to be the first line of defense. However, according to a survey by Ponemon Institute, 69% organizations do not believe that AVs can block the threats that they monitor. Given these staggering numbers, classical rule-based malware detectors can simply not be expected to detect the large influx of malware variants. The main problem with rule-based defenses is that they are reactive, where a rule is added only *after* experiencing an attack.

Machine Learning (ML) has become a promising ally for malware detection. The security community has been investigating ways to incorporate machine learning for intelligent malware detection, profiling, and analysis. Figure 1 shows a typical pipeline for malware defense and the opportunities to introduce machine learning in it. It is noteworthy that machine learning is also useful for attackers: Due to the intrinsic adversarial nature of the threat landscape, machine learning has not only been used to build intelligent defenses, but it has also been used to develop intelligent attacks that evade detection. In the past decade alone, this arms-race has resulted in more than 20,000 research articles.

[1] https://www.av-test.org/en/statistics/malware/.
[2] https://www.accenture.com/us-en/insight-cost-of-cybercrime-2017.

© Springer Nature Switzerland AG 2022
L. Batina et al. (Eds.): Security and Artificial Intelligence, LNCS 13049, pp. 217–253, 2022.
https://doi.org/10.1007/978-3-030-98795-4_10

In this chapter, we conduct a systematic survey of the literature published in the past decade to establish a taxonomy of the main research themes. For an unbiased literature review, we select peer-reviewed papers containing a combination of fixed search queries that are highly cited in their domain. We summarize the state-of-the-art in various sub-fields of intelligent malware defenses, *i.e.,* malware detection, malware analysis, adversarial malware, and malware author attribution. The literature is greatly dominated by *malware detection* approaches with the aim of developing scalable behavioral signatures. Approaches from other domains have been applied to perform malware detection, such as natural language processing, image visualization, graph mining, and bioinformatics. We categorize the research in this domain according to the data source and feature representation used for their classifiers. *Malware analysis* is another research direction that develops tools that provide the necessary insights to improve malware detection. We discuss approaches that aim to increase interpretability, and provide smarter ways to collect behavioral traces. *Adversarial machine learning* has recently gained popularity, not only for machine learning-based offensive security, but also for hardening machine learning classifiers. Finally, *malware author attribution* aims to associate malware to its author(s), a field that is mainly driven by law enforcement agencies. Although not a very active area, it serves as a powerful use-case for interdisciplinary research. Figure 2 shows the literature overview in a chronological order, divided across the aforementioned research directions.

We discuss important considerations that emerge when machine learning is applied to malware, such as resilience against concept drift and evasion, handling imbalanced datasets, using appropriate evaluation metrics, and providing privacy and performance guarantees. We have observed that the absence of toy problems, representative datasets, explainable approaches, and the usage of noisy ground truth has limited the reproducibility of available research. Specifically, explainable approaches are necessary for debugging existing techniques and developing newer ones based on obtained insights. These issues need to be addressed by the research community in order to encourage systematized research in the intelligent malware defenses domain.

This chapter is organized as follows. Section 2 serves as a roadmap for the rest of the chapter: it identifies the feature sources and representations that have been used to characterize malware in the literature, including several feature engineering modes. Section 3 discusses the vast literature that explores effective and efficient malware detection methods. We expand the discussion on malware research in Sect. 4 by covering relevant areas, *i.e.,* malware analysis, adversarial malware, and author attribution. Section 5 enumerates the main challenges unique to machine learning-based malware defenses. Section 6 highlights the four key issues that should be addressed to enable reproducible research in the intelligent malware defenses domain. Finally, we conclude our discussion in Sect. 7.

2 Malware Characterization

The success of machine learning classifiers lies in finding the data that appropriately characterizes malware. Determining these data and the input features

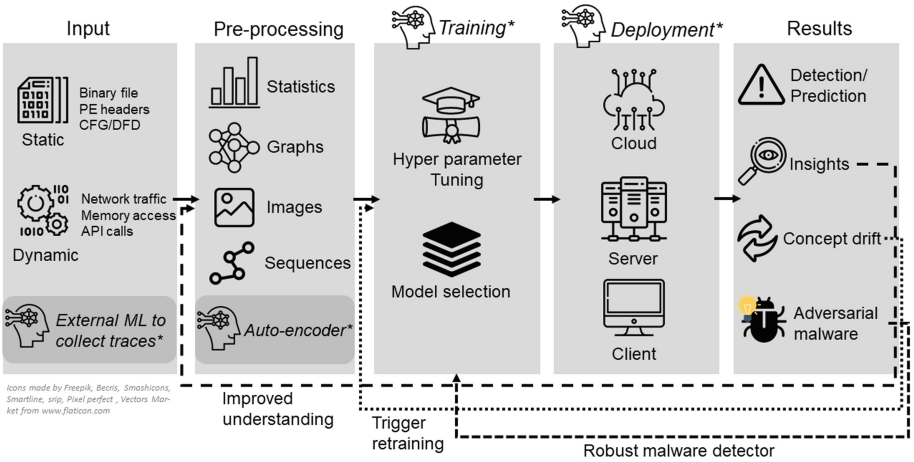

Fig. 1. Machine learning pipeline for malware defense: Raw input data gets transformed into features during pre-processing, which are given as input for model training. The raw data may be collected using ML-based sandboxing. Features may be automatically learned using auto-encoders. Depending on the problem, the model may be deployed at different locations, and the output may be used in several ways: Malware detection typically raises alerts; Insights obtained from malware analysis can lead to improved malware characterization; Detecting changes in data distribution can trigger model retraining; Incorporation of adversarial malware in the training process can lead to robust models.

required for machine learning is a difficult task since they will be used to detect new malware samples that may behave in unexpected ways. Anderson *et al.* [10] suggest that effective feature engineering, including features obtained from domain experts, plays a key role in classifier performance. There exists a myriad of literature exploring the various features that can be used to characterize different types of malware [122,132]. In this chapter, we provide an overview of the past ten years of intelligent malware defenses from a technical machine learning perspective, grounded in the types of features used to characterize malware: statistical, graphs, images, and sequences. The type of input feature greatly influences which machine learning technology can be employed. Other important considerations are: (a) the target platform of the malware, (b) how to collect data from a malware, and (c) how to extract features from such data. We briefly introduce these other considerations below but will not go into detail.

2.1 Platform-Specific Malware and Defenses

Malware often targets consumer devices, like desktop computers and handheld devices. The first malware was a PC-based virus, called Elk Cloner[3],

[3] https://en.wikipedia.org/wiki/Elk_Cloner.

discovered over 35 years ago. Since then, malware has targeted multiple operating system and browser vulnerabilities to infect desktop computers. It is important to keep in mind that most research in intelligent malware defense also targets specific platforms, and this has a great effect on the used input features. Recent ransomware [4] and cryptojacking malware [30] are well-known for attacking desktops. A major strain in research therefore targets desktop malware, most frequently Windows-based malware. Recently, Ucci *et al.* [122] presented a survey of features characterizing Windows Portable Executable (PE) malware.

With the widespread use of Android smartphones, there is a growing amount of literature on Android malware detection [54,66,72,83,104,118,119,134,144,145], mainly for two reasons: Firstly, Android is an open-source operating system, so developers can investigate various vulnerabilities that malware has exploited over the years; Secondly, multiple large-scale and open-source datasets of malicious Android applications have supported research in this field. This does not mean that iOS-based malware does not exist [87], it is just not a frequent subject of research.

Nowadays consumer devices are increasingly becoming equipped with Internet connectivity, known as the *Internet of Things (IoT)*. This comes with new risks culminating in a novel strain of malware specifically targeting IoT devices [131]. IoT devices are commonly made available on the Internet with their default configurations, which makes them an easy target for the attackers. Since IoT devices have limited computational resources, their built-in security is significantly inferior to other internet-connected devices, making it an ideal use-case for intelligent malware defenses. Several recent works have proposed to enhance IoT security using ML-based malware detection [13,40]. In this chapter, we try to avoid this distinction between platforms, instead focusing on technological differences. However, often the platform and technologies are tightly linked, and solutions are typically not directly applicable to other platforms.

2.2 Feature Sources

There are two major approaches for analyzing malware: (i) static analysis, and (ii) dynamic analysis. Machine learning has been applied successfully in both approaches. For static analysis, *static features* are extracted from a malware's code, *i.e.,* without executing it. For dynamic analysis, features are extracted by running malware and monitoring its behavior. These features can be obtained from two sources: standard *dynamic features* are generated on the host device, typically by interacting with the operating system, while *network features* are created from network traffic generated by hosts in a network.

Static Features. The source code of a malware, often obtained by decompiling its binary, is the most reliable artefact to identify its objective. Early studies on malware detection have mainly characterized malware using features extracted from its code. These features are often obtained by doing a comparative analysis of goodware (benign software) and malware features, and selecting the ones that

are observed more frequently in malware. PE headers are commonly used for Windows malware [3,33,47,59,107], while some works extract features directly from the malware binary itself [56,60,96]. Most of the recent literature from static analysis comes from the smartphone domain (*i.e.,* Android), where the features are extracted from either the APK's manifest file or the disassembled Dalvik bytecode for signature generation. Shabtai *et al.* [111] is one of the earlier works on Android malware detection that uses features from the Dalvik bytecode in order to perform binary classification (*i.e.,* malware vs. goodware). One of their major contributions is to apply a myriad of classification algorithms and select the one that achieves the highest accuracy. Other features, such as Operational Codes (OpCodes) [40,72] and function API calls [1,5,54,105,134], are also commonly used for malware detection. Existing works also analyze the amount and types of permissions that applications request to measure their maliciousness [63,104,134,144]. For a higher-level semantic analysis, features from Control flow graphs [104,137] and Data flow graphs [39] have also been used.

Dynamic Features. With the widespread use of code obfuscation tools to evade detection and to generate malware variants, syntactic analysis has become increasingly more difficult. Additionally, there has been a spike in fileless malware infections[4], where the malware code resides purely in the victim host's memory without leaving any code fingerprints. Hence, dynamic analysis is more popular for malware detection. In dynamic analysis, malware is executed in a controlled environment and its behavior is monitored [94]. Information such as, system API calls [19,34,75,141], memory access patterns [55,74], and running processes [20] are common sources for feature selection. Some works consider inter-file relationships between files present on a system for malware detection [95,120]. There also exist hybrid approaches, where static analysis guides dynamic analysis for thorough code coverage [117].

Network Features. Network traffic analysis is popular because it can be performed remotely and presents lower overhead than its system-activity counterpart. However, machine learning has been slow to materialize in the network security domain because of noisy ground truth and non-stationary data distribution [10]. Existing sandboxes also have limited support for handling network requests due to the risk of lateral movement, *i.e.,* when the attacker who has gained access to the network spreads their reach to other hosts [45]. Nevertheless, the use of HTTP header fields [27,66,93] and traffic connections [83] is common for malicious traffic detection. Privacy concerns have also been addressed in network security. Boukhtouta *et al.* [18] evaluate the differences between Deep Packet Inspection (DPI) based methods and IP-header based methods for classifying malicious network traffic. They conclude that IP-header features make the machine learning model generalizable and can achieve higher accuracy due to the independence from packet payloads. They also suggest that using IP-header features can help fingerprint zero-day malware, *i.e.,* malware never seen before.

[4] https://www.cybereason.com/blog/fileless-malware.

2.3 Feature Engineering Modes

Most of the existing works perform manual feature engineering, where the features are initially obtained from domain experts and are further cherry-picked based on a classifier's accuracy [1,5,33,40,54,59,72,93,96,105,107,134]. The downside to manually engineering features is that it is a laborious and (potentially) subjective process, which may need to be repeated in case of concept drift, for example. The malware threats also need to be fully scoped-out before arriving at the optimal feature-set. Nevertheless, knowing the features beforehand helps with explaining and debugging a classifier's decisions.

In the recent years, deep neural networks have gained tremendous traction, making automated feature engineering through representation learning (*e.g.,* with auto-encoders) a popular choice. Pascanu *et al.* [89] propose a recurrent neural network approach, specifically Echo State Networks, for automatically extracting time-domain features. They use these features in a Logistic Regression (LR) model for malware classification, and achieve better performance than a trigram based manual feature engineering approach. David *et al.* [28] propose a system that uses Deep Belief Networks (DBN), which are a type of generative graphical deep neural network that can perform unsupervised learning, with a deep stack of de-noising auto-encoders to automatically generate behavioral signatures. Yuxin *et al.* [137] have also used DBNs as auto-encoders to automatically extract features from malware executables.

A common critique faced by features that are automatically learned from deep neural networks is that they are uninterpretable, and hence undesirable for building explainable solutions. Building interpretable deep learning models is an open area of research. Zhu *et al.* [145] have recently proposed an interesting approach that automatically engineers features by mimicking human analysts' feature engineering processes. Their system mines academic documents and synthesizes their knowledge into interpretable features that are later used for Android malware classification. They report comparable results to state-of-the-art manual feature engineering approaches.

2.4 Feature Representation

After having selected the data source and features to use, the next step is to determine how to process these data. We identify four different kinds of feature representations in this chapter, *i.e.,* statistical, graphs, images, and sequences. Note that feature representation means the format of the input given to a ML classifier, not the intermediate representations from representation learning.

Represent as Statistics. The most common representation for features in the literature is statistical [1,3,5,10,16,20,34,63,66,73,77,93,97,99,104–107,119,134,144]. Feature values are collapsed using aggregates or correlations. Some statistical features also capture the temporal aspect of the behavior without having to deal with sequential data, *e.g.,* the Power Spectral Density of the FFT of the Command & Control communication [121].

Represent as Graphs. One branch of literature represents malware using graphs, *e.g.,* to represent the hosts a malware connects with, as an abstraction over the original feature-set to get a high-level view of malware's behavior. Graphs are used to either ease analysis [139] or to extract semantic features for malware detection [32,42,95,120]. Graphs are also used to perform malware causality analysis [68,70,140].

Represent as Images. Another branch of literature explores various visualization methods in order to characterize malware. A malware binary can be converted into an image by encoding the raw values of the binary as the color intensity of pixels [2,50,56,84,114,130]. The intuition here is that a malware family may share similar code pieces, which will appear as similar motifs in the image. Visualizing malware, and hence exposing these similarities to the human eye, can potentially aid manual malware analysis. Furthermore, research has also applied standard image classification techniques to perform efficient and reasonably accurate malware detection. Despite these encouraging initial results, we note that image representation for malware classification should be considered with caution, as discussed in more detail in Sect. 3.3.

Represent as Sequences. There is an increasing amount of literature that is investigating the use of sequential data in behavior characterization. Although sequential features capture the temporal aspect of behavior, machine learning algorithms with inherent support for sequences are rare. The main difficulty lies in appropriately measuring the distance between two sequences in the presence of noise, delays and misalignments. Methods from other fields, such as *sequence alignment* adopted from bioinformatics [26,57], and *n-grams* adopted from natural language processing [22,35,40,47,52] have been utilized to that end. Increasingly more approaches are using deep neural networks because they have good support for sequences, *e.g.,* Long Short Term Memory networks (LSTM) [75,141], Recurrent Neural Networks (RNN) [43,59,112] and Word2Vec [9,21,54].

Fig. 2. Overview of ML-based malware defenses proposed over the past decade.

Table 1. Literature according to feature source and representation.

Representation	Source		
	Static	Dynamic	Network
Statistical	[1,5,63,73,77,104–107,119,134,144]	[19,20,34,77,99]	[10,16,66,77,93,97,121]
Graphs	[32,42,104]	[68,70,95,120]	[140]
Images	[2,50,56,60,84,114,130]	–	[125]
Sequences	[13,21,22,26,28,33,40,47,54,59,72,73,89,96,137,145]	[57,75,141]	[36,91]

3 Malware Detection

A central objective of malware research is to develop behavioral signatures that can automatically detect future malware variants. We make a distinction between two major strains of research in the malware domain: *Detection-based* (this section) and *Analysis-based* (Sect. 4.1).

Most of the existing literature is about malware detection and signature generation, with the end-goal of optimizing metrics, such as classifier accuracy and F1-scores [16,36,77,91,93,97,121]. To this end, a plethora of research has been conducted over the past ten years that explores various features, representations, and machine learning algorithms. LeDoux *et al.* [61] summarize the research on malware detection, particularly on code-reuse detection, using machine learning. They also enumerate malware analysis problems that machine learning *is* equipped to solve. The vast literature on feature extraction and data mining techniques for malware detection is comprehensively described by Ye *et al.* [132]. Souri *et al.* [116] evaluate various signature- and behavior-based intelligent malware detectors.

We survey malware detection methods by focusing on the feature representations used by the applied machine learning or data mining technology. Both supervised and unsupervised machine learning techniques have been used in the literature. We categorize them into three classes: (i) *Binary classification*: Determine whether an unlabeled software is goodware or malware, (ii) *Multi-class classification*: Given a set of unlabeled malware samples and a set of known malware family names, perform malware family attribution, and (iii) *Clustering*: Given a set of unlabelled software, categorize them into distinct classes based on structural/behavioral differences. Table 1 summarizes the malware detectors reviewed in this chapter, categorized according to the feature source (*i.e.,* static, dynamic, network) and input representation (*i.e.,* statistical, graph, image, sequence) that they employ.

3.1 Statistical Approaches

An aggregated feature-set is the most widespread feature representation used in literature. Statistical features are fast to compute and simple to incorporate in a machine learning classifier.

Experiments have shown that the prevalence of certain feature values is a decent indicator of malware. Naturally, binary classification proves to be more

successful under this setting than multi-class classification. Hence, even if the experimental dataset is composed of malware from multiple families, earlier works considered them together as one 'malicious' class. Alazab *et al.* [5] perform binary classification for the task of 'zero-day' malware detection. They use the frequency of static API calls to characterize Windows binaries, and show that it is indeed possible to distinguish between malware and goodware using this characterization. Aafer *et al.* [1] propose a light-weight Android malware detector based on the frequency of static API calls. They use K-Nearest Neighbors (KNN) to detect and alert the user of malicious applications. Sahs *et al.* [104] characterize Android applications using a binary vector of used permissions. Because there exist significantly more benign applications than malicious ones, they utilize a one-class Support Vector Machine (SVM) which characterizes benign applications well, and helps detect Android malware. Yerima *et al.* [134] use Bayesian classification to detect malware. They characterize applications using API call- and permission-frequencies.

Santos *et al.* [106] utilize the frequency of OpCode occurrence in executables to detect malware. They show that Polynomial Kernel classifiers and Random Forests achieve the best performance, which is not surprising because these algorithms have a long history of performing well in text classification. Suarez *et al.* [119] extract statistical features from Control Flow Graph (CFG) code blocks, such as the number and redundancy of code chunks, and common and discriminant code chunks. These text-based features are used for multi-class classification. Earlier deep learning approaches have also characterized malware using aggregates. For example, Saxe *et al.* [107] propose a malware detection system that uses a feature-set of byte histograms and frequency of PE import calls.

Although statistical features have been applied successfully in a supervised manner, there are assumptions that may be difficult to realize in practice. For example, AV-assigned family labels are noisy and novel threats are common. In contrast, unsupervised learning can be used independently to identify novel threats. In the binary classification setting, an anomaly detection approach is often used to model the benign class and anomalies are labelled as malicious, giving the capability to detect novel threats. For example, Burguera *et al.* [19] built a malware detection system using K-Means clustering to identify anomalous system events by finding deviations from one 'normal' cluster. For the multi-class setting, clustering is used to identify different threat classes. Perdisci *et al.* [93] present one of the first unsupervised clustering approaches to detect HTTP-based malware. They propose multi-step clustering to enable large-scale malware behavioral signature generation. Unsupervised machine learning is also often used in combination with supervised approaches to improve detection capabilities. Rieck *et al.* [99] propose an incremental analysis approach for malware family identification: by first performing clustering to identify *novel malware classes*, and then classifying unknown malware samples by assigning them to these discovered classes. Burnap *et al.* [20] have recently developed an unsupervised learning method based on Self Organizing Feature Maps (SOMs) that cluster similar malware behavior. They use the clusters of similar behavior as features for later classification tasks. The key benefit of this approach is an added

layer of abstraction for improved classification—instead of using raw features for classification, the system allows fuzzy boundaries that can map new samples onto the existing decision boundary. David *et al.* [28] use Deep Belief Networks, a type of unsupervised model, to automatically generate malware behavioral signatures. Finally, Li *et al.* [66] build a network traffic-based malware classifier that utilizes both supervised and unsupervised classifiers to improve classification accuracy.

Effective malware detectors provide stable and trustworthy results. Since statistical features provide an aggregated view, a malicious application may appear similar to a benign application from this point of view. Hence, the choice of feature representation plays a crucial role in a classifier's robustness. Recently, Milosevic *et al.* [73] compared two text-mining approaches for Android malware detection—using statistical features (*i.e.,* permissions) and sequential features (*i.e.,* bag-of-words of decompiled `Dex` code). Their experiments show that the bag-of-words approach performs better due to better malware characterization, indicating that statistical features may not be the optimal choice in all cases. Another way to improve the robustness of malware detectors is to use *Ensemble learning*, *i.e.,* a learning paradigm that combines the decisions of multiple classifiers to arrive at the final decision. Ensemble models, such as Random Forests have been shown to be robust to non-stationary data distribution, such as network traffic [10]. Recently, Zhu *et al.* [144] proposed an ensemble Rotation Forests model to classify Android malware. Rotation Forests [100] are an ensemble of Decision Trees where diversity through rotated principal components is given emphasis, resulting in more stable decisions.

3.2 Graph-Mining Approaches

Graph-mining approaches have been used to represent malware's relationships in a graphical format, in order to provide an added abstraction layer. Malware literature adopts scalable graph mining approaches to perform fast detection.

Security products often have to mark files as malicious or benign based on a partial view of the host's file system. Approaches using inter-file relationships have emerged as a solution. Chau *et al.* [95] perform malware detection using large-scale graph inference. They consider files as malware based on *guilt-by-association*—they exploit the inter-file relationships present on multiple systems to compute reputation scores for unlabelled files. They use Belief Propagation algorithm to mark files with low reputation as malware. They evaluate their approach on a 60 terabyte dataset composed of a Billion-node graph, and show significant improvement over existing approaches. Acar *et al.* [120] also use a similar approach to detect malware, with the additional use of Locality Sensitive Hashing (LSH) for efficiently binning similar files.

Hou *et al.* [42] build the first approach that uses a Structured Heterogeneous Information Network (HIN) to characterize API-relatedness. The HIN edges are used to measure the semantic similarity between API calls, that is used to measure maliciousness of an application using multi-kernel learning. Fan *et al.* [32] improve upon the previous approach using a *meta-graph* to determine inter-file relationships for malware detection. For cost-effectiveness, they use *MetaGraph2Vec* to learn low-dimensional representations for the HIN that

preserves its structure and semantics. *MetaGraph2Vec* is a meta-graph approach that has shown competitive performance for heterogeneous graph-mining tasks such as node classification and clustering.

Using graphs to perform causality analysis is also an area of interest. Zhang *et al.* [140] detect malware by performing *causality analysis* on its network traffic. They build Triggering Relation Graphs (TRG) that show the inter-dependency of various network events. The TRGs show an absence of dependency between legitimate and malicious network events, hence making it easier to detect malicious activities. Liu *et al.* [68] build a backward- and forward-causality graph to detect abnormal system events, based on their rareness and location in the causality graph.

3.3 Image Visualization Approaches

Malware visualization has opened a new research direction that uses ML-based image classification to detect malware. These approaches rely on converting a malware binary into an image which is then provided to an image classifier, either as a raw image or as a set of extracted features. The key assumptions for these methods are: (a) *malware families have similar images because of code reuse*, and (b) *malware images are significantly different from goodware images.*

In 2011, Nataraj *et al.* [84] proposed a straightforward method to convert malware into an image: a malware executable represented as a binary vector is reshaped into a matrix of an arbitrary width and is viewed as a grey-scale image. The authors observed that malware binaries belonging to the same family were visually similar in both layout and texture. They extracted textural features from these images and applied K-Nearest-Neighbors to perform malware family identification. This approach proved to be very efficient and achieved high accuracy, close to prior methods that used static features such as n-grams. Furthermore, they showed that malware belonging to one family packed with the same packer, or containing sections encrypted with polymorphic engines, are still categorized together as the same family, indicating some level of resilience to *naïve* obfuscation. Motivated with the initial positive results, several follow-up studies [2,56] expanded the research by investigating different types of image feature extractors and machine learning classifiers. Approaches to malware classification based on image similarity were confirmed to be effective on the commonly used Kaggle Microsoft Malware dataset [101].

Recent work has started applying deep neural networks for the classification task, inspired by their encouraging performance in the image classification domain. Kalash *et al.* [50] successfully perform malware family identification with a two-dimensional Convolutional Neural Network (CNN) architecture, and Singh *et al.* [114] convert malware binaries into colored-images to classify obfuscated malware with a `ResNet-50` architecture, *i.e.,* a CNN architecture with shortcut connections providing superior performance. Yakura *et al.* [130] applied a CNN with attention mechanism [128] that allows to explain which areas of an image contribute to particular classification decisions. Le *et al.* [60] have

recently proposed a fully automated malware classification approach for non-domain experts. They represent raw binary files as grey-scale images using the approach by Nataraj *et al.* [84]. The images are given to a hybrid network, *i.e.,* CNN with Bi-directional LSTM model, which outperforms a traditional CNN model.

Despite promising performance on benchmark datasets, representing binaries as images for malware detection comes with several limitations. Apart from some empirical results, there appears to be little evidence to strongly support the aforementioned underlying assumptions of malware classification based on visual similarity of binaries. Some studies have encountered images from different malware families that exhibited such similar patterns that they were classified in one class [2]. Naturally, only previously seen malware can be identified, while zero-day malware which is structurally different will likely evade detection [56]. Even for known malware, a common case of false negatives is observed when malicious content of relatively small size is embedded within a goodware, such that the resulting image remains very similar to other benign examples. Additionally, when global image features are used, merely relocating sections in a binary or adding large sections with redundant data may be sufficient to alter the image texture and mislead a classifier. These issues illustrate that generalizability of this approach, beyond particular datasets, remains an open question.

While image representations allow large-scale malware classification in a computationally feasible way, deep neural networks do not necessarily require this intermediate transformation. One-dimensional CNN architectures and sequential deep neural networks are directly applicable to raw one-dimensional binaries. Meanwhile, reshaping an initially one-dimensional binary sequence to a two-dimensional image with an arbitrarily chosen width introduces artificial spatial relations that are not present in the original file. Additionally, this representation does not appear stable, as adding or removing binary data in one location may completely change the positional relations in the vertical direction of the converted image [130]. Such artefacts may obscure naturally occurring patterns in data and negatively impact classification. This line of reasoning may also apply to a few studies that attempt to use CNN on dynamic features, such as malicious network traffic: Wang *et al.* [125] perform malware detection by representing network traffic as images. In any case, it is very challenging to make such two-dimensional CNNs capable of recognizing the temporal dependencies required for processing traffic data.

3.4 Sequence Learning Approaches

Sequential pattern mining has emerged as a promising approach for malware detection due to increasingly better-performing sequential ML classifiers.

Earlier sequence learning works utilize n-grams to characterize temporal behavior. Jain *et al.* [47] represent PE files as n-grams (with varying values for n) to perform binary classification. They select the prominent n-grams using Class-wise Document Frequency method. Their experiments show that trigrams with Random Forests give the best malware detection performance. Similarly,

Canfora *et al.* [22] use n-grams generated from OpCode sequences to detect malware families. They show that bigrams with Random Forests give the best performance. We believe these differences exist due to different datasets and feature selection approaches. Fan *et al.* [33] uses function call sequences and an All-Nearest Neighbors (ANN) classifier for PE-malware detection. ANN is a modification of K-Nearest Neighbors algorithm that makes a decision based on *all* neighbors. ANN makes a compromise on run-time efficiency for the sake of robustness.

In recent works, deep learning-based approaches have dominated malware detection using sequence learning. Convolutional Neural Networks (CNN) are used due to their ability to detect complex and non-linear patterns in data. Raff *et al.* [96] build a CNN framework that takes an entire PE binary as input for automated feature engineering and malware detection. Their method consumes the entire executable as opposed to only the PE-header to avoid over-fitting on the header features. However, their results do not show significant improvement over their baseline—a byte n-gram model. Mclaughlin *et al.* [72] perform Android malware detection using a CNN framework that utilizes raw OpCode sequences. Azmoodeh *et al.* [13] also use OpCode sequences to perform binary classification on IoT malware that is specifically used for military purposes. They reduce the feature dimensions using Principle Component Analysis (PCA) by providing only the first two components to the classifier. Recently, Cakir *et al.* [21] have used Word2Vec feature embedding on OpCode sequences to characterize malware. They use Gradient Boosting, which is a type of ensemble learner, for binary classification of malware. Karbab *et al.* [54] use CNN and Word2Vec feature embedding on API call sequences for malware family identification. They evaluate their system on multiple datasets, such as the MalGenome dataset [143], the Drebin dataset [12], and benign applications from Google Play[5]. Kolosnjaji *et al.* [59] perform multi-class classification using hybrid deep learning, *i.e.,* Convolutional and Recurrent Neural Networks (CNN/RNN), for Windows malware family identification.

Recently, Haddadpajouh *et al.* [40] have explored various configurations of Long Short Term Memory networks (LSTMs) for IoT malware classification. They characterize binaries by their OpCode sequences, and then choose the features that maximize the Information Gain (IG). However, it is unclear how generalizable their results are as they only use ~500 binaries for the classification task. Zhang *et al.* [141] use so-called behavior chains based on API call sequences to characterize malware's behavior. They use LSTMs to perform binary classification and report a false positive rate of less than 2% (in the best case). Mishra *et al.* [75] use deep learning and the sequence of dynamic system calls for malware classification in the cloud environment. Their system uses two layered approach – CNN for feature engineering and Bi-directional LSTMs for malware detection. They evaluate their system on a university's network traffic and show promising results.

Sequential features are a common occurrence in bioinformatics. As computer viruses attain their characterization from similarity to natural viruses,

[5] https://play.google.com/store/apps.

bioinformatics-inspired solutions have also been proposed to detect malware variants. A common approach to detect similar DNA sequences is through the use of sequence alignment algorithms, such as the Smith-Waterman algorithm. Sequence alignment methods work by assigning a score based on matches, mismatches and gaps. These values are embedded in a substitution matrix. Domain-specific substitution matrices exist for bioinformatics applications. Chen *et al.* [26] map malware's binary code to Amino acid characters, and use the so-called Residue substitution matrix, while Naidu *et al.* [82] report that the `PAM-350` substitution matrix performs the best for malware variant detection. Chen *et al.* [26] also develop a multiple sequence alignment method that uses neural networks to classify viruses and worms. They show that alignment-based methods allow classifiers to find similarities with more ease compared to other methods.

3.5 Performance Optimizations

Malware detectors need to be efficient to cope with the exponential increase in malware attacks. Hence, some works propose extensions to existing works for improving the performance of traditional malware detectors.

Feature Reduction. A classifier's performance is directly dependent on the quality of features used for model learning. The key idea is to select the least number of features that maximally characterize a malware. Hence, a straightforward optimization is to conduct a feature reduction step that eliminates redundant features. Li *et al.* [63] develop a fast Android malware detector using an SVM classifier. As a feature reduction step, they perform 'significant permissions' analysis, that selects only the permissions that distinguish between malicious and benign applications with high confidence. Their results achieve up to 32 times speed-up compared to two competing approaches, i.e. Drebin [12] and Permission-induced Risk Malware Detection [124]. Similarly, Yerima *et al.* [134] select only the features with maximum Mutual Information (MI) to speed up their malware detector. Firdausi *et al.* [34] demonstrate elevated performance of their malware detector after conducting a best-first feature selection process.

Hardware-Assisted Detection. Hardware-assisted Malware Detection (HMD) has emerged as an alternative for improving malware detection using Hardware Performance Counters (HPC). HMDs are light-weight detectors that live on the microprocessor to provide a first-line of defense, and to reduce overhead on software-based detectors. Khasawneh *et al.* [55] show that hardware-detectors reduce performance overhead by up to 11 times compared to software-only detectors. Xu *et al.* [129] propose a novel HMD that monitors system calls' memory access patterns, which are used to classify malware, *e.g.,* kernel rootkits.

Existing studies suggest that HMDs execute a malware sample multiple times in order to collect the required data, due to the limited number of HPCs available on microprocessors. Most of the methods propose to use an ensemble of light-weight classifiers to resolve this issue. Khasawneh *et al.* [55] propose an ensemble of *specialized* LR classifiers to improve the performance of HMDs,

while only inducing minimal additional overhead on the cycle-time and power consumption. *Specialized classifiers* are malware-family specific classifiers, *i.e.,* one classifier is trained for one class of malware. They use LR due to its cheap and simple implementation on the microprocessor. Sayadi *et al.* [110] propose ensemble learning to collect the required data while using even less HPCs. Their results show that ensemble learning approach using only 4 HPCs can match the robustness and performance of standard classifiers that use 16 HPCs. In a recent work, Sayadi *et al.* [109] demonstrate that the performance of HMDs is directly related to the number of available HPCs. They propose a feature reduction step in order to select the most significant HPCs. They propose a two-step classification approach: a *course-grained classifier* that categorizes a software as either goodware or one of the malicious classes (*i.e.,* rootkit or trojan); followed by a *fine-grained specialized classifier* (*i.e.,* one for each type of malicious class). To further reduce the run-time, they utilize ensemble learning in the coarse-grained classifiers and show that using merely 4 HPCs outperforms state-of-the-art classifiers with 8 HPCs by a factor of 1.31.

3.6 Trend

There is a growing interest towards alternative approaches for malware detection, such as causality-based, and ant-colony optimization-based approaches [44], and a more targeted focus on Android malware detection. The use of sequence learning is growing, especially due to superior performance of recurrent neural networks, such as LSTMs, but also due to sequences being better equipped to characterize behavior. For example, Amer *et al.* [9] have used a combination of Word2Vec and Markov chains to establish the relationship between malware API call sequences. Despite that, works on network-traffic based sequential models are meagre due to the difficulty of handling non-stationary and noisy sequences. There is also a growing concern for the brittle nature of neural networks, to be discussed in Sect. 4.2, which is driving research towards better interpretability of such models' output.

4 Additional Research Directions

Although malware detection is a central research objective, there are additional research directions that have been gaining traction lately. *Malware analysis*, as opposed to detection, aims to improve malware understandability rather than to optimize detection rates [17,78,119]. *Adversarial machine learning* techniques have gained particular popularity in recent years in relation to detecting evasive malware. Finally, *attributing malware* to its author(s) is also an area of interest, mainly driven by law enforcement agencies. In this section, we discuss the seminal works in these three popular research themes.

4.1 Malware Analysis

Malware analysis methods aim to improve malware understandability, and provide essential insights that can improve malware detection methods. Although

malware analysis and detection are often seen together in literature, below we present approaches that: (a) aim to improve malware understandability, *e.g.,* by providing insights into malware relationships, and (b) enable malware analysis, *e.g.,* by collecting traces and building analysis environments.

Ucci *et al.* [122] present a recent survey on ML-based malware analysis techniques. The survey provides a taxonomy of research objectives, features and ML algorithms used for Windows PE malware. They identify topical trends on malware triage. They also present the concept of *Malware Analysis Economics* that studies the trade-off between detection accuracy and the resources required for detection.

Increasing Interpretability. Malware analysts have to frequently monitor large-scale network traffic, which is a laborious task. Zhang *et al.* [139] propose a framework to visualise the causal relationships between network requests to help detect abnormal events. Their user studies reveal that visualising network traffic in this way enhances analysts' malware detection capabilities. Mariconti *et al.* [70] perform causality analysis on user actions that trigger a malware infection. They characterize malware samples by the trigger-actions commonly performed by users. Their method can successfully infer relations between, *e.g.,* information-stealing malware and web pages asking for user credentials. Suarez *et al.* [119] build a dendrogram of malware families showing overlapping code snippets, which helps them to generate evolution-invariant signatures.

Smith *et al.* [115] have pointed towards the semantic gap between the machine learning and malware analysis communities. One of their proposals is to reposition the task from *identifying malware* to *identifying behavior*, making it possible to understand what a malware is doing. Along these lines, Nadeem *et al.* [81] have proposed the use of behavioral profiles to describe malware samples as opposed to using black-box family names. They develop MalPaCA, a clustering-based framework that discovers distinct behaviors present in network traffic and uses the cluster membership information to generate a profile for each malware sample.

Collecting Traces. Collecting malware traces, especially for dynamic analysis, is a challenging problem due to the difficulty of finding live malware samples and setting up sandboxes. Burguera *et al.* [19] addresses the unavailability of malware datasets by setting up a crowd-sourcing system to collect system traces from unlimited number of real smartphone users. Secondly, effective features for malware are often not shared among the security community. Gu *et al.* [39] address this issue by introducing a consortium blockchain framework. The blockchain is used as a database of malware-characterizing features. Their classifier consumes the blockchain for malware family identification. Recently, Shibahara *et al.* [112] have proposed a machine learning-based data collection method for efficient dynamic analysis. Typically, malware traces are collected for a fixed amount of time before moving on to executing the next sample. The method proposed in [112] treats network traces as natural language, and uses the communication

pattern as a heuristic to suspend analysis. They use RNN to learn the underlying objective of communication and to detect when a change in purpose occurs. They suspend analysis when a malware has stopped its activities. With this approach, they report a reduction of 67.1% analysis time.

Sandboxing. Multiple works have proposed sandboxes that can forcefully trigger malware functionality in order to provide more holistic behavioral logs, but very few have used machine learning to do so. There also exist other approaches that use search-based algorithms to improve code coverage of malware samples, *e.g.,* Wang *et al.* [126] propose a fuzzing-based approach to forcefully trigger malware's hidden behaviors. Among the literature that uses machine learning is the work by Spreitzenbarth *et al.* [117]. They propose an end-to-end analysis environment for Android malware where applications are executed, traces are collected and a clustering algorithm categorizes them as malware or goodware. However, they only use machine learning to *post-process* behavioral traces. Additionally, their sandbox does not support latest Android versions. Yerima *et al.* [133] have recently proposed a machine learning based malware analysis framework. They learn a state machine of each Android application using code's static analysis. They use insights from the state machine to guide the so-called *stateful event generation.* They also compare with an existing approach based on random event-generation and show that the guided behavior-triggering approach results in better data collection.

An orthogonal research objective is *sandbox evasion*, where a malware uses machine learning to detect whether it is being executed in a sandbox or on a live system. When malware detects the presence of a sandbox, it either shuts down, or starts sending garbage data to mislead analysis. Yokoyama *et al.* [135] show that it is possible for attackers to use straightforward machine learning algorithms to differentiate between a sandbox and a live system based on leaking characteristics of Windows-based sandboxes. Miramirkhani *et al.* [74] propose sandbox evasion techniques based on the natural 'wear and tear' of a real system compared to that of a sandbox. They exploit the past usage of a system to determine its age and degree of use. They show that a simple decision tree classifier can differentiate between a sandbox and a real system with a very high accuracy.

Trend. The security community appears to be heavily biased towards detection-based solutions. Analysis is most often conducted as a precursor for detection methods, or as a part of Systematization of Knowledge studies. Recently, there has been a push towards using explainable machine learning for the malware domain, which specifically allows to reason about classifier decisions. Fan *et al.* [31] have evaluated various explanation techniques for malware analysis, and conclude that LIME [98] and SHAP [69] provide the most robust and stable explanations. Also, machine learning has not yet been applied to malware lineage—how a certain malware family evolves over time in terms of structure, behavior and its target. Most of the work in this domain is manual and requires a deep understanding of the evolving threat landscape: *e.g.,* Black *et al.* [17] per-

form an in-depth analysis of banking malware families, and Moubarak *et al.* [78] discuss the structural relationship between several potentially state-sponsored malware. Evidently, the success of this research is dependent on the quality and size of the used dataset.

4.2 Adversarial Malware

The security of machine learning is an active area of research that has been gaining increasing popularity in the recent years. This theme addresses the arms-race between crafting evasive malware samples (offensive security) and developing robust methods to detect said samples (defensive security). Although the threat landscape is already adversarial in nature, many approaches in this area have been borrowed from the computer vision domain, where adversarial ML was pioneered. Biggio *et al.* [15] provide an overview of the developments in adversarial machine learning in the past ten years. An open problem in this area is using machine learning to craft adversarial malware samples where the perturbations are big enough to mimic goodware while preserving malicious functionality.

Offensive Security. While evasive malware has existed for a long time, latest research applies machine learning to automatically craft these samples. These techniques work by performing small perturbations on a malware sample to create a variant that leads to a misclassification by the ML model. Most of the proposed attacks are gradient-based, as they target deep neural networks. White-box techniques require some knowledge of the target, such as the structure and weights of the target model, while black-box techniques do not assume any knowledge of the targeted classifier.

There are two main concerns in creating adversarial malware samples: (a) the perturbations are performed in the continuous domain, while malware binaries exist in the discrete domain; and (b) the frameworks often create perturbations that break functionality of the executable. Anderson *et al.* [11] have proposed a reinforcement learning-based method to guide the search for functionality-preserving perturbations. However, since their method is quite general, they report modest evasion rates. Grosse *et al.* [38] propose a method for crafting adversarial examples that operates in the discrete domain and preserves functionality. They craft adversarial Android malware by adding constraints to the perturbations—they only allow changes in the manifest file that adds a single line of code to the application. They use the adversarial examples on Drebin [12] and report a misclassification (evasion) rate of 69%. Hu *et al.* [43] target RNN models based on sequential API features. They learn a local substitute (surrogate) model of the victim RNN that propagates the gradients to a generative RNN that produces sequential adversarial examples. Their results show that more than 90% of the adversarial examples result in misclassifications. Kolosnjaji *et al.* [58] target a sequential model that learns from raw malware bytes. They craft adversarial examples using a gradient-based attack modifying the last 1% of the bytes that achieve misclassifications. They report a maximum evasion rate

of 60%. Al-Dujaili *et al.* [3] adapt the saddle-point optimization problem from the continuous domain to generate adversarial examples in the discrete domain. They present a framework that discovers adversarial examples and incorporates them in the training process to harden the learnt classifier. They conclude that the randomized rounding technique helps discover four times as many adversarial examples. Chen *et al.* [25] craft adversarial Android malware by optimally perturbing the Dalvik byte code to target semantic features. They show the effectiveness of their adversarial examples by using them on two famous Android malware detectors, MaMaDroid [86] and Drebin [12], where they report an evasion rate of 100%. Recently, Verwer *et al.* [123] used [3] to develop GRAMS, which is a greedy approach that randomly flips bits to obtain functionality-preserving high-quality adversarial examples in the discrete domain. GRAMS was successful in crafting evasive malware and defending against competitors' evasion attempts during the robust malware detection challenge[6].

Poisoning attacks are another important concern for machine learning classifiers. Poisoning attacks refer to an attacker's capability to inject adversarial examples during classifier *training phase*, such that it learns to classify malicious entities as benign. For example, Biggio *et al.* [14] poison behavioral malware clustering and Muñoz-González *et al.* [79] propose a poisoning algorithm for deep learning classifiers. Chen *et al.* [24] have specified attacker models for poisoning attacks in the malware domain: (a) a *weak attacker* who injects malicious code in the non-logical part of the application, such as manifest file; (b) a *strong attacker* who injects malicious code in resources, such as `jar` or `jpg`; and (c) a *sophisticated attacker* who uses Dynamic Code Loading via Reflection for injecting malicious code at run-time. Having concrete attacker models provides terminology to develop more streamlined defenses, and to compute resilience guarantees.

Defensive Security. One approach for forensic malware analysis is to categorize malware based on similar evasion strategies. Kirat *et al.* [57] propose a bioinformatics-inspired solution to generate and analyze evasion signatures—they cluster similar evasive behavior among malware samples. They use a sequence alignment algorithm to measure similarity among different system call sequences. Then, they extract evasion signatures from the behavioral clusters. These signatures can be used to detect when a future malware sample attempts to evade detection in a similar way.

One of the key benefits of adversarial machine learning is that it hardens the security of an adversarially trained model [38]. When adversarial examples are part of the training process, they allow to discover samples in the so-called *blind spots* of the malicious domain, increasing its robustness to unseen evasive samples. Recently, there has been a lot of interest in developing ML-based adversary-aware approaches. The main difference from malware detectors previously discussed in Sect. 3 is that these approaches actively anticipate evasion attempts. Demontis *et al.* [29] propose a so-called *secure-learning* paradigm that

[6] https://github.com/ALFA-group/malware_challenge.

suggests having the feature weights more evenly distributed in order to bound linear classifiers' sensitivity to feature changes. They also propose attacker models based on their capabilities, knowledge and skills. Zhang *et al.* [138] propose an adversary-aware feature selection method, since the choice of features may also be a factor in adversarial robustness. Their wrapper-based framework makes assumptions about the adversary and simulates evasion attacks at each step of the training phase. The framework chooses the features that maximize the classifier's generalizability in the absence of adversarial examples and minimize the classifier's impact against evasion attempts. Chen *et al.* [23] present a robust malware detection system based on two key components: (a) a *feature selection method* that picks features that maximize attacker's evasion costs, and (b) an *ensemble learning method* with diverse classifiers that incorporate a major part of the feature space. Li *et al.* [62] investigate the resilience of ensemble classifiers and the effectiveness of ensemble attacks. Their experiments show that while adversarial training for ensemble classifiers promotes robustness, they are unfortunately no match for adversarial examples learned through ensemble methods.

Another way to harden classifiers without explicit adversarial training is through the special handling of suspicious files. Chen *et al.* [24] have developed a self-adaptive learning scheme for detecting poisoning attacks. They introduce a so-called *camouflage detector* that finds suspicious false negatives by performing similarity analysis with the most-benign and most-malicious looking samples, and sends such *camouflaged samples* back to the training phase as malicious examples.

Trend. At the moment of writing, the security community seems to have a strong affinity towards offensive security research. Naturally, conducting defensive research is especially challenging due to the strict requirements that a defensive framework is expected to fulfill. In case of malware defenses, provable robustness to evasion is a major milestone that cannot be reached by the community without public datasets of evasive malware. We believe that adversarial learning for defensive model hardening is an unfolding but promising research field.

4.3 Malware Author Attribution

In general, authorship attribution can be considered from two perspectives: (i) *code authorship attribution* – attributing a software to its author(s); and (ii) *family attribution* – identifying code similarities between unlabelled software pieces. In malware research, the latter problem appears as *malware family identification*, which involves multi-class classification already covered in Sect. 3. Hence, here we discuss the code authorship attribution problem.

Code authorship attribution has a rich history in the Software Engineering literature. The aim of this research is to extract features that capture an author's programming style. Existing work can also be found in related fields, such as forgery and plagiarism detection where the goal is to extract distinguishing stylistic or fingerprinting attributes from a software that identifies where the

code was copied from. Source code attribution is the simplest variant since the author's stylistic features can be relatively easily extracted. One such work is proposed by Alsulami *et al.* [8] who extract features from the Abstract Syntax Tree (AST) of source code collected from Google Code Jam (GCJ). In fact, GCJ and Github are popular sources of experimental data for authorship attribution, in general [6,7,103,113].

In many real-world settings however, source code is not directly available, rendering aforementioned techniques ineffective. It is also commonly believed that the compilation process removes most of the stylistic features. Rosenblum *et al.* [103] perform one of the first attempts to address this problem by using a ML approach that identifies the surviving stylistic features for binary authorship attribution.

It is noteworthy that code authorship attribution for malware is significantly more difficult because the authors have strong incentives to hide their identity. Using machine learning to solve the attribution problem is also tricky because code samples from known malware authors that are required to train a classifier are rarely available. Additionally, the availability of malware-as-a-service indicates that samples are authored by multiple developers in a malware's lifetime. Hence, research in this area is scarce because of the difficulty of establishing a ground truth.

Saxe *et al.* [108] have written an introductory book on big data analysis for malware detection. They show the usage of static and dynamic analysis for performing *shared code analysis* with the aim of identifying similar adversary groups. Alrabaee *et al.* [6] propose a multi-layered approach to improve malware binary authorship attribution by conducting both syntax- and semantic- analysis. They attempt to reconstruct the source code from malware binaries, which they compare with code-based signatures of other families. They also extract semantic features, such as the way registers are manipulated, to establish strong evidence for attribution. Rosenberg *et al.* [102] use deep neural networks for the attribution of nation-state Advanced Persistent Threats (APTs). They observe that nation-state APTs have different styles and objectives, which makes their classification feasible. To that end, they take raw dynamic logs as input to the neural network that learns a high-level abstraction of the APTs. Their system is evaluated on two major nation-state malware families and show it to be effective for the purpose.

Natural Language Processing has also been proposed for attribution purposes after its success in attributing cyber-stalkers [35]. Kalgutkar *et al.* [52] build an Android malware detection system based on 'malware author' signatures. Their system leverages strings extracted from the malware binaries to generate profiles of malware authors, with the expectation that future malware samples authored by the same person will match the profiles. They use n-grams for feature representation and Support Vector Machine for APK classification.

Research related to adversarial attacks also exists in code authorship attribution. Simko *et al.* [113] propose adversarial stylometry attacks to defeat source code attribution classifiers. They demonstrate that current code attribution clas-

sifiers are not robust to adversarial attacks, even when they are executed by non-experts. The authors claim that although not fool-proof, augmenting machine learning classifiers with human analysts proves to be more resilient against adversarial attacks, especially when they are warned about potential forgeries in the code. They analyze C/C++ programs and conclude that semantic features, such as those extracted from ASTs are more resilient to forgery attacks.

Alrabaee *et al.* [7] present a literature survey of existing techniques for malware binary attribution. The survey also lists features that can be used for author attribution because they survive compilation, *e.g.,* compiler information, system calls, and the usage of particular strings may characterize coding styles. Additionally, certain type of bugs in the code may also point to semantic hints that can be used for author attribution. They also note that a key research challenge is feature selection that captures author's style rather than functionality of the program. Following this, Murenin *et al.* [80] have used LIME to understand the role of selected features in source code attribution for Android malware.

Iqbal *et al.* [46] have recently published a book on authorship attribution and cyber forensics using machine learning in which they comprehensively describe research into authorship identification and attribution using few training samples, authorship characterization and verification. Kalgutkar *et al.* [51] show how the field has evolved from basic software matching techniques to sophisticated methods based on API calls and dependency graphs. They conclude that although there is no one-size-fits-all solution yet for malware attribution, the existing work on varying levels of abstraction has brought us one step closer to the solution. Nevertheless, this field still has many open problems that are yet to be explored.

Trend. The popularity of the malware attribution field is impacted by the adversarial and ad-hoc nature of the threat landscape. Unavailability of open datasets further complicates realistic evaluation. Further, having a narrow target audience (*e.g.,* law enforcement, intelligence agencies) means that the works do not get highly cited and thus remain undiscovered. We believe that this field can get a new life with explainable approaches, open benchmark datasets and access to ground truth.

5 Challenges in ML-Applied Malware Defenses

The malware domain presents unique challenges for machine learning application. After years of research, the security community has made a significant headway in highlighting the proper usage of machine learning for malware defenses. As a result, additional problems have emerged that require further investigation. Souri *et al.* [116] and Ye *et al.* [132] identify several unsolved problems in the data-mining based malware detection domain. One must remember that machine learning is not a silver bullet that can solve all malware-related problems [51]. In this section, we describe common pitfalls and challenges that emerge when ML

is used for malware detection, which should be accounted for when designing and evaluating such methods.

Robustness Against Time-Decay. Some of the existing work is filled with unrealistic simplifying assumptions about the malware landscape. One of the most prevalent assumptions is the *closed-world assumption*, which assumes that training data is fully representative of all categories of samples that may appear at test-time. However, as malware is an ever-evolving threat, static training data will inevitably become outdated. Consequently, researchers have shown that ML-classifiers' performance degrades over time [92,107]. Recent works have incorporated *concept drift detection* in their ML classifiers for handling non-stationary data population. These classifiers continually re-learn the changing concepts in order to maintain an acceptable detection accuracy. Jordaney *et al.* [48] and Wang *et al.* [127] use P-values that can proactively detect concept drift before the classifier's performance starts to degrade. There is also a growing interest in semantic features that are less affected by malware evolution and hence slow down the aging of malware detectors [142].

Robustness Against Evasion. Evasion resilience is an important characteristic for deployable classifiers. A misleading expectation from ML classifiers is that they should be *fully and provably evasion resilient*. To this end, defensive adversarial machine learning has emerged as a promising solution for evasion resilient classifiers, which has been previously discussed in Sect. 4.2. The purpose of defensive adversarial ML is to explore additional search space in order to harden models against evasion attempts. However, this search is still bounded by the Independent and Identically Distributed data (i.i.d) assumption. As a consequence, out-of-distribution adversarial examples prevalent in the open world are unlikely to be detected by an adversarially trained ML classifier.

Imbalanced Training-Set. Benign examples occur significantly more frequently than malicious ones. Failure to incorporate this trait in the training dataset creates a so-called *spatial-bias* [92] in the classifier. Existing works have often used unrealistic class distribution, *e.g.,* the use of inverted class distribution [83] and equal class distribution [90,136]. Chen *et al.* [27] propose a solution for imbalanced network traffic classification to perform accurate Android malware detection. They experiment with various combinations of imbalanced classification algorithms, such as Synthetic Minority Oversampling Technique (SMOTE) with SVM, SVM cost-sensitive and `C4.5` cost-sensitive. They also develop Simplex Imbalanced Data Gravitation-based Classification (S-IDGC) that works faster while maintaining the stability of IDGC. In the deep learning domain, Le *et al.* [60] use the class re-balance sampling procedure in bi-directional LSTMs to address the class imbalance problem.

Evaluation Metrics. The usage of appropriate evaluation metrics is an underrated challenge. For example, using accuracy to measure a classifier's performance when it is trained with a highly imbalanced dataset results in misleading conclusions. Similarly, precision and recall values can also be altered based

on the choice of empirical thresholds and dataset-specific parameters. Meanwhile, evaluating approaches using such varied metrics limits objective analysis, and the obtained results become incomparable. Jordaney *et al.* [49] demonstrate that traditional evaluation metrics show misleading information about classifiers' performance. They propose two metrics based on non-conformity measures for evaluating a classifiers' performance. *Credibility* measures the homogeneity of a given label compared to others of the same class, and *Confidence* measures the separation between a given label and other classes. Pendlebury *et al.* [92] have recently identified experimental biases in existing Android malware classifiers, namely (a) *spatial bias* due to unrealistic class distribution in training and testing data, and (b) *temporal bias* due to incorrect time splits causing impossible configurations. They propose a new metric, namely Area Under Time (AUT), to characterize classifier robustness when time decay is present.

Privacy Concerns. Machine learning classifiers typically perform better with fine-grained contextual features. In an attempt to perform large-scale classification, classifiers have access to both benign and malicious data. Privacy concerns arise as the feature-set becomes more and more fine-grained. For example, DPI-based approaches analyze the payload of each packet, which may contain privacy sensitive information. With data protection laws being widely enforced, such methods are tricky to deploy at large-scale. There are a couple of solutions for being privacy-aware: (i) selecting abstract features that do not violate the privacy of user actions, while still being able to characterize malicious behavior; (ii) deploying a distributed classifier, as in the case of federated learning, that trains on local data provided by multiple clients [85]. In the latter solution, secure multi-party computation (SMC) and differential privacy (DP) are required to provide privacy guarantees.

Performance Optimizations. Malware infections are a widespread security threat faced by all network-connected devices. Classical machine learning solutions are often not ideal for fending off millions of malware infections each day. Effective intelligent defenses should be fast, proactive and evolve with the changing threat landscape. Hence, a dedicated research direction exists that designs online, optimized classifiers capable of detecting malware in real-time. Federated learning, discussed earlier as a possible solution to privacy issues, provides a distributed infrastructure that enables efficient large-scale detection. Other works are discussed in Sect. 3.5.

6 Open Problems in ML-Based Malware Defenses

In this section, we discuss what we believe are the key problems that the research community should address. At the heart lies the problem of reproducible research: the absence of *toy problems* and *representative datasets* makes the results from different papers incomparable. Furthermore, the results often cannot be taken at their face value because malware *ground truth* is inherently inconsistent and unreliable. Crucially, many solutions eagerly emphasize metric optimization but overlook *explainability*, providing little new insight into

the problem of malware detection. We believe that these four issues hinder fair assessment of new contributions in the intelligent malware defenses domain. It is very difficult to objectively compare new methods against state-of-the-art solutions for the same problem, using the same data, and the same ground truth. An alarming side-effect is the lack of meaningful contributions to the field even though many new papers are published each year.

Toy Problems. Toy problems are important in the early development of a research field. These are simplified challenges that can help develop and test methodologies that solve a more challenging problem. Computer science in general and artificial intelligence in particular have established traditional toy problems that are still used to develop newer methodologies. However, malware research has always aimed to solve real-world threats. We observe that limited access to data and resources that are necessary for the evaluation of proposed methodologies has constrained systematic and open academic research. Moreover, building fool-proof methodologies in malware detection is an especially challenging problem, because the adversaries keep evolving rapidly. Since malware is constantly evolving, the research is driven by the availability of newer threats, and is reactive in nature. In light of these inevitable issues, we urge the community to introduce standardized toy problems, which could act as a starting point for developing new methods in a more synchronized and proactive manner. Toy problems would allow the assessment of proposed algorithms in isolation from the general practical limitations of malware detection. Ultimately, as the algorithms become more mature, they should be enhanced for deployment—practical feasibility should not be fully discarded at the envisioning stage.

Representative Datasets. The biggest hurdle in ML-based malware analysis research is the absence of representative datasets. These datasets are crucial for the development of usable and generalizable defensive solutions. However, with the rapid evolution of malware, any available dataset becomes obsolete in a matter of years, *e.g.,* the well-known VX Heavens dataset [41] from 2010, the Drebin dataset [12] from 2010–2012, and the MalGenome dataset [143] from 2012 are arguably no longer representative. Since most of the available datasets are not representative, the trained models only describe part of the real threat landscape. This is not to say that open-source datasets are not available. In fact, the Stratosphere IPS project[7] has published large-scale network traffic, *e.g.,* the CTU-13 dataset [37] captures traffic for 13 botnet scenarios, and the recently published IoT-23 dataset [88] captures 20 IoT malware scenarios and 3 benign ones. The Kaggle Microsoft Malware dataset [101] was also widely used in multiple works. Other works have also released their private datasets for reproducibility [67,105]. Nevertheless, one promising way for the academic community to gain access to reliable and representative data is to establish a long-term collaboration with industry partners who directly monitor the threat landscape and can provide updated threat intelligence for the development of robust machine learning solutions. This is an excellent way to keep up with

[7] https://www.stratosphereips.org/datasets-overview.

the rapidly changing threat landscape. The downside is that the data often contains highly sensitive information that cannot be released to the public, thus exacerbating problems of reproducibility.

Noisy Ground Truth. Existing literature has repeatedly shown that AV-provided malware family labels are inconsistent. These labels are used as ground truth by researchers to evaluate newer malware detection methods, making the results unreliable. Popular tools, such as VirusTotal[8], run multiple AV scanners and return an array of labels predicted by each scanner, without any indication as to which is correct. There is also an absence of a common vocabulary that all security companies can follow to label malware samples. Research has shown that the consensus reached by AV scanners regarding the labeling of a single malware sample is no better than a coin-toss (around 50%) [76]. Machine learning has also been slow to materialize in network security domain due to non-stationary data and noisy ground truth [10]. Unsupervised ML can already provide a foundation to address this issue. However, in practice, existing unsupervised ML approaches often use some form of ground truth for evaluation. For example, Perdisci *et al.* [93] evaluate their malware clustering by introducing a notion of AV graphs that depict the agreement between AV vendors as a measure of cluster cohesion and separation. Yuping *et al.* [65] use majority-voted family labels from 25 AV vendors as their ground truth to evaluate malware clustering. Li *et al.* [64] have advised caution when deciphering highly accurate clustering results as they can be impacted by spatial bias: performing majority voting on AV-provided labels is hazardous, since if most of the AV vendors are in agreement, it typically indicates that the families are already easy to detect. Hence, we either need better ground truth [53] or purely data-driven unsupervised evaluation approaches.

Explainable Solutions. In recent works, deep learning based malware detectors have surpassed the performance of traditional ML classifiers. They have also automated the detection pipeline for the most part. However, deep neural networks are inherent black-boxes that provide limited interpretability. It is also alarming how brittle deep learning is to adversarial attacks. Alternatively, non-deep learning approaches are not much more interpretable—they are frequently packed with complicated filtering steps to maximise performance [77], which also turns them into black-boxes. This concern has motivated research on explainable machine learning. Explainable models enable identification of bias in raw data, debug errors in trained models, enable model optimization, and allow analysts to extrapolate advanced results, *i.e.,* to get detailed insights from data instead of simply reading off detection rates. Without explainability, such extrapolation will be difficult and error-prone. Mathews *et al.* [71] provide a summary of explainable ML techniques for malware classification, including both intrinsic and post-hoc methods. The research trend shows that moving forward, special emphasis will be given to explainable and human-in-the-loop solutions.

[8] https://www.virustotal.com/.

7 Summary

Machine learning has emerged as a promising ally for developing intelligent malware defenses. However, the research in this area is scattered across different venues and domains. In this chapter, we identify the key research themes and assemble the state-of-the-art literature that has been proposed in the past decade. In doing so, we highlight trends in these research themes.

The literature is greatly dominated by *malware detection* approaches with the aim of developing scalable behavioral signatures. We categorize the research in this domain according to the data source and feature representation used for their classifiers. The trends in the literature suggest that sequence learning and explainable machine learning are considered promising areas of research. *Malware analysis* is another research direction that develops tools that provide the necessary insights to improve malware detection. *Adversarial machine learning* has recently gained popularity to harden machine learning classifiers. Also, *malware author attribution* proves to be a challenging field with limited progress due to the unavailability of datasets, and an absence of concrete problem statements that data-driven methods can realistically address.

We have discussed important considerations that emerge when machine learning is applied in the malware domain, such as resilience against concept drift and evasion, handling imbalanced datasets and using appropriate evaluation metrics. We have also identified key issues that need to be addressed in our opinion by the research community in order to encourage systematized research in the malware domain: toy problems, representative datasets, noisy ground truth, and explainable solutions. Without overcoming these issues, limited progress can be made due to the inability to compare research results.

It is evident that intelligent malware defenses will continue to grow. However, understanding the unique challenges that the malware domain brings to the table is absolutely essential for developing effective machine learning enabled solutions that can withstand the test of time.

References

1. Aafer, Y., Du, W., Yin, H.: DroidAPIMiner: mining API-level features for robust malware detection in Android. In: Zia, T., Zomaya, A., Varadharajan, V., Mao, M. (eds.) SecureComm 2013. LNICST, vol. 127, pp. 86–103. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-04283-1_6
2. Ahmadi, M., Ulyanov, D., Semenov, S., Trofimov, M., Giacinto, G.: Novel feature extraction, selection and fusion for effective malware family classification. In: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, pp. 183–194 (2016)
3. Al-Dujaili, A., Huang, A., Hemberg, E., O'Reilly, U.M.: Adversarial deep learning for robust detection of binary encoded malware. In: 2018 IEEE Security and Privacy Workshops (SPW), pp. 76–82. IEEE (2018)
4. Al-rimy, B.A.S., Maarof, M.A., Shaid, S.Z.M.: Ransomware threat success factors, taxonomy, and countermeasures: a survey and research directions. Comput. Secur. **74**, 144–166 (2018)

5. Alazab, M., Venkatraman, S., Watters, P., Alazab, M., et al.: Zero-day malware detection based on supervised learning algorithms of API call signatures. In: 2011 Australasian Data Mining Conference (AusDM 11) (2010)
6. Alrabaee, S., Saleem, N., Preda, S., Wang, L., Debbabi, M.: OBA2: an onion approach to binary code authorship attribution. Digit. Investig. **11**, S94–S103 (2014)
7. Alrabaee, S., Shirani, P., Debbabi, M., Wang, L.: On the feasibility of malware authorship attribution. In: Cuppens, F., Wang, L., Cuppens-Boulahia, N., Tawbi, N., Garcia-Alfaro, J. (eds.) FPS 2016. LNCS, vol. 10128, pp. 256–272. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51966-1_17
8. Alsulami, B., Dauber, E., Harang, R., Mancoridis, S., Greenstadt, R.: Source code authorship attribution using long short-term memory based networks. In: Foley, S.N., Gollmann, D., Snekkenes, E. (eds.) ESORICS 2017. LNCS, vol. 10492, pp. 65–82. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66402-6_6
9. Amer, E., Zelinka, I.: A dynamic windows malware detection and prediction method based on contextual understanding of API call sequence. Comput. Secur. **92**, 101760 (2020)
10. Anderson, B., McGrew, D.: Machine learning for encrypted malware traffic classification: accounting for noisy labels and non-stationarity. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1723–1732 (2017)
11. Anderson, H.S., Kharkar, A., Filar, B., Roth, P.: Evading machine learning malware detection. Black Hat (2017)
12. Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K., Siemens, C.: Drebin: effective and explainable detection of Android malware in your pocket. In: NDSS, vol. 14, pp. 23–26 (2014)
13. Azmoodeh, A., Dehghantanha, A., Choo, K.K.R.: Robust malware detection for internet of (battlefield) things devices using deep eigenspace learning. IEEE Trans. Sustain. Comput. **4**(1), 88–95 (2018)
14. Biggio, B., et al.: Poisoning behavioral malware clustering. In: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, pp. 27–36 (2014)
15. Biggio, B., Roli, F.: Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recogn. **84**, 317–331 (2018)
16. Bilge, L., Balzarotti, D., Robertson, W., Kirda, E., Kruegel, C.: Disclosure: detecting botnet command and control servers through large-scale netflow analysis. In: ACSAC, pp. 129–138. ACM (2012)
17. Black, P., Gondal, I., Layton, R.: A survey of similarities in banking malware behaviours. Comput. Secur. **77**, 756–772 (2017)
18. Boukhtouta, A., Mokhov, S.A., Lakhdari, N.E., Debbabi, M., Paquet, J.: Network malware classification comparison using dpi and flow packet headers. J. Comput. Virol. Hacking Tech. **12**(2), 69–100 (2016)
19. Burguera, I., Zurutuza, U., Nadjm-Tehrani, S.: CrowDroid: behavior-based malware detection system for Android. In: Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, pp. 15–26 (2011)
20. Burnap, P., French, R., Turner, F., Jones, K.: Malware classification using self organising feature maps and machine activity data. Comput. Secur. **73**, 399–410 (2018)
21. Cakir, B., Dogdu, E.: Malware classification using deep learning methods. In: Proceedings of the ACMSE 2018 Conference, pp. 1–5 (2018)

22. Canfora, G., De Lorenzo, A., Medvet, E., Mercaldo, F., Visaggio, C.A.: Effectiveness of opcode ngrams for detection of multi family Android malware. In: 2015 10th International Conference on Availability, Reliability and Security, pp. 333–340. IEEE (2015)
23. Chen, L., Hou, S., Ye, Y.: SecureDroid: enhancing security of machine learning-based detection against adversarial Android malware attacks. In: Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 362–372 (2017)
24. Chen, S., et al.: Automated poisoning attacks and defenses in malware detection systems: an adversarial machine learning approach. Comput. Secur. **73**, 326–344 (2018)
25. Chen, X., et al.: Android HIV: a study of repackaging malware for evading machine-learning detection. IEEE Trans. Inf. Forensics Secur. **15**, 987–1001 (2019)
26. Chen, Y., Narayanan, A., Pang, S., Tao, B.: Malicioius software detection using multiple sequence alignment and data mining. In: 2012 IEEE 26th International Conference on Advanced Information Networking and Applications, pp. 8–14. IEEE (2012)
27. Chen, Z., et al.: Machine learning based mobile malware detection using highly imbalanced network traffic. Inf. Sci. **433**, 346–364 (2018)
28. David, O.E., Netanyahu, N.S.: Deepsign: deep learning for automatic malware signature generation and classification. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2015)
29. Demontis, A., et al.: Yes, machine learning can be more secure! a case study on Android malware detection. IEEE Trans. Dependable Secure Comput. **16**, 711–724 (2017)
30. Eskandari, S., Leoutsarakos, A., Mursch, T., Clark, J.: A first look at browser-based cryptojacking. In: 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 58–66. IEEE (2018)
31. Fan, M., Wei, W., Xie, X., Liu, Y., Guan, X., Liu, T.: Can we trust your explanations? Sanity checks for interpreters in Android malware analysis. IEEE Trans. Inf. Forensics Secur. **16**, 838–853 (2020)
32. Fan, Y., Hou, S., Zhang, Y., Ye, Y., Abdulhayoglu, M.: Gotcha-sly malware! scorpion a metagraph2vec based malware detection system. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 253–262 (2018)
33. Fan, Y., Ye, Y., Chen, L.: Malicious sequential pattern mining for automatic malware detection. Expert Syst. Appl. **52**, 16–25 (2016)
34. Firdausi, I., Erwin, A., Nugroho, A.S., et al.: Analysis of machine learning techniques used in behavior-based malware detection. In: 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp. 201–203. IEEE (2010)
35. Frommholz, I., Al-Khateeb, H.M., Potthast, M., Ghasem, Z., Shukla, M., Short, E.: On textual analysis and machine learning for cyberstalking detection. Datenbank-Spektrum **16**(2), 127–135 (2016)
36. Garcia, S.: Modelling the network behaviour of malware to block malicious patterns. The stratosphere project: a behavioural IPS. Virus Bulletin (2015)
37. Garcia, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. Comput. Secur. **45**, 100–123 (2014)
38. Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P.: Adversarial examples for malware detection. In: Foley, S.N., Gollmann, D., Snekkenes, E. (eds.) ESORICS 2017. LNCS, vol. 10493, pp. 62–79. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66399-9_4

39. Gu, J., Sun, B., Du, X., Wang, J., Zhuang, Y., Wang, Z.: Consortium blockchain-based malware detection in mobile devices. IEEE Access **6**, 12118–12128 (2018)
40. HaddadPajouh, H., Dehghantanha, A., Khayami, R., Choo, K.K.R.: A deep recurrent neural network based approach for internet of things malware threat hunting. Futur. Gener. Comput. Syst. **85**, 88–96 (2018)
41. VX Heaven: VX heaven virus collection, 15 May 2010. http://vxheaven.org/
42. Hou, S., Ye, Y., Song, Y., Abdulhayoglu, M.: HinDroid: an intelligent Android malware detection system based on structured heterogeneous information network. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1507–1515 (2017)
43. Hu, W., Tan, Y.: Black-box attacks against RNN based malware detection algorithms. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
44. Huang, H., Deng, H., Sheng, Y., Ye, X.: Accelerating convolutional neural network-based malware traffic detection through ant-colony clustering. J. Intell. Fuzzy Syst. (Preprint) **37**, 1–15 (2019)
45. Ijaz, M., Durad, M.H., Ismail, M.: Static and dynamic malware analysis using machine learning. In: 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 687–691. IEEE (2019)
46. Iqbal, F., Debbabi, M., Fung, B.C.: Machine Learning for Authorship Attribution and Cyber Forensics. Springer, Heidelberg (2020)
47. Jain, S., Meena, Y.K.: Byte level n–gram analysis for malware detection. In: Venugopal, K.R., Patnaik, L.M. (eds.) ICIP 2011. CCIS, vol. 157, pp. 51–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22786-8_6
48. Jordaney, R., et al.: Transcend: detecting concept drift in malware classification models. In: 26th USENIX Security Symposium (USENIX Security 2017), pp. 625–642 (2017)
49. Jordaney, R., Wang, Z., Papini, D., Nouretdinov, I., Cavallaro, L.: Misleading metrics: on evaluating machine learning for malware with confidence. Technical report (2016)
50. Kalash, M., Rochan, M., Mohammed, N., Bruce, N.D., Wang, Y., Iqbal, F.: Malware classification with deep convolutional neural networks. In: 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. IEEE (2018)
51. Kalgutkar, V., Kaur, R., Gonzalez, H., Stakhanova, N., Matyukhina, A.: Code authorship attribution: methods and challenges. ACM Comput. Surv. (CSUR) **52**(1), 1–36 (2019)
52. Kalgutkar, V., Stakhanova, N., Cook, P., Matyukhina, A.: Android authorship attribution through string analysis. In: Proceedings of the 13th International Conference on Availability, Reliability and Security, pp. 1–10 (2018)
53. Kantchelian, A., et al.: Better malware ground truth: techniques for weighting anti-virus vendor labels. In: Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, pp. 45–56 (2015)
54. Karbab, E.B., Debbabi, M., Derhab, A., Mouheb, D.: MalDozer: automatic framework for Android malware detection using deep learning. Digit. Investig. **24**, S48–S59 (2018)
55. Khasawneh, K.N., Ozsoy, M., Donovick, C., Abu-Ghazaleh, N., Ponomarev, D.: Ensemble learning for low-level hardware-supported malware detection. In: Bos, H., Monrose, F., Blanc, G. (eds.) RAID 2015. LNCS, vol. 9404, pp. 3–25. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26362-5_1

56. Kirat, D., Nataraj, L., Vigna, G., Manjunath, B.: SigMal: a static signal processing based malware triage. In: Proceedings of the 29th Annual Computer Security Applications Conference, pp. 89–98 (2013)
57. Kirat, D., Vigna, G.: MalGene: automatic extraction of malware analysis evasion signature. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 769–780 (2015)
58. Kolosnjaji, B., et al.: Adversarial malware binaries: evading deep learning for malware detection in executables. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 533–537. IEEE (2018)
59. Kolosnjaji, B., Zarras, A., Webster, G., Eckert, C.: Deep learning for classification of malware system call sequences. In: Kang, B.H., Bai, Q. (eds.) AI 2016. LNCS (LNAI), vol. 9992, pp. 137–149. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50127-7_11
60. Le, Q., Boydell, O., Mac Namee, B., Scanlon, M.: Deep learning at the shallow end: malware classification for non-domain experts. Digit. Investig. **26**, S118–S126 (2018)
61. LeDoux, C., Lakhotia, A.: Malware and machine learning. In: Yager, R.R., Reformat, M.Z., Alajlan, N. (eds.) Intelligent Methods for Cyber Warfare. SCI, vol. 563, pp. 1–42. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-08624-8_1
62. Li, D., Li, Q.: Adversarial deep ensemble: evasion attacks and defenses for malware detection. IEEE Trans. Inf. Forensics Secur. **15**, 3886–3900 (2020)
63. Li, J., Sun, L., Yan, Q., Li, Z., Srisa-an, W., Ye, H.: Significant permission identification for machine-learning-based Android malware detection. IEEE Trans. Ind. Inf. **14**(7), 3216–3225 (2018)
64. Li, P., Liu, L., Gao, D., Reiter, M.K.: On challenges in evaluating malware clustering. In: Jha, S., Sommer, R., Kreibich, C. (eds.) RAID 2010. LNCS, vol. 6307, pp. 238–255. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15512-3_13
65. Li, Y., Jang, J., Hu, X., Ou, X.: Android malware clustering through malicious payload mining. In: Dacier, M., Bailey, M., Polychronakis, M., Antonakakis, M. (eds.) RAID 2017. LNCS, vol. 10453, pp. 192–214. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66332-6_9
66. Li, Z., Sun, L., Yan, Q., Srisa-an, W., Chen, Z.: DroidClassifier: efficient adaptive mining of application-layer header for classifying Android malware. In: Deng, R., Weng, J., Ren, K., Yegneswaran, V. (eds.) SecureComm 2016. LNICST, vol. 198, pp. 597–616. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59608-2_33
67. Lindorfer, M., Neugschwandtner, M., Weichselbaum, L., Fratantonio, Y., Van Der Veen, V., Platzer, C.: Andrubis-1,000,000 apps later: a view on current Android malware behaviors. In: 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), pp. 3–17. IEEE (2014)
68. Liu, Y., et al.: Towards a timely causality analysis for enterprise security. In: NDSS (2018)
69. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774 (2017)
70. Mariconti, E., Onaolapo, J., Ross, G., Stringhini, G.: The cause of all evils: assessing causality between user actions and malware activity. In: 10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 2017) (2017)

71. Mathews, S.M.: Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In: Arai, K., Bhatia, R., Kapoor, S. (eds.) CompCom 2019. AISC, vol. 998, pp. 1269–1292. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22868-2_90
72. McLaughlin, N., et al.: Deep Android malware detection. In: Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, pp. 301–308 (2017)
73. Milosevic, N., Dehghantanha, A., Choo, K.K.R.: Machine learning aided Android malware classification. Comput. Electr. Eng. **61**, 266–274 (2017)
74. Miramirkhani, N., Appini, M.P., Nikiforakis, N., Polychronakis, M.: Spotless sandboxes: evading malware analysis systems using wear-and-tear artifacts. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 1009–1024. IEEE (2017)
75. Mishra, P., Khurana, K., Gupta, S., Sharma, M.K.: VMAnalyzer: malware semantic analysis using integrated CNN and bi-directional LSTM for detecting VM-level attacks in cloud. In: 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1–6. IEEE (2019)
76. Mohaisen, A., Alrawi, O., Larson, M., McPherson, D.: Towards a methodical evaluation of antivirus scans and labels. In: Kim, Y., Lee, H., Perrig, A. (eds.) WISA 2013. LNCS, vol. 8267, pp. 231–241. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-05149-9_15
77. Mohaisen, A., Alrawi, O., Mohaisen, M.: AMAL: high-fidelity, behavior-based automated malware analysis and classification. Comput. Secur. **52**, 251–266 (2015)
78. Moubarak, J., Chamoun, M., Filiol, E.: Comparative study of recent MEA malware phylogeny. In: 2017 2nd International Conference on Computer and Communication Systems (ICCCS), pp. 16–20. IEEE (2017)
79. Muñoz-González, L., et al.: Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 27–38 (2017)
80. Murenin, I., Novikova, E., Ushakov, R., Kholod, I.: Explaining Android application authorship attribution based on source code analysis. In: Murenin, I., Novikova, E., Ushakov, R., Kholod, I. (eds.) NEW2AN/ruSMART -2020. LNCS, vol. 12525, pp. 43–56. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65726-0_5
81. Nadeem, A., Hammerschmidt, C., Gañán, C.H., Verwer, S.: Beyond labeling: using clustering to build network behavioral profiles of malware families. In: Stamp, M., Alazab, M., Shalaginov, A. (eds.) Malware Analysis Using Artificial Intelligence and Deep Learning, pp. 381–409. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-62582-5_15
82. Naidu, V., Narayanan, A.: Using different substitution matrices in a string-matching technique for identifying viral polymorphic malware variants. In: 2016 IEEE Congress on Evolutionary Computation (CEC), pp. 2903–2910. IEEE (2016)
83. Narudin, F.A., Feizollah, A., Anuar, N.B., Gani, A.: Evaluation of machine learning classifiers for mobile malware detection. Soft. Comput. **20**(1), 343–357 (2014). https://doi.org/10.1007/s00500-014-1511-6
84. Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S.: Malware images: visualization and automatic classification. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security, pp. 1–7 (2011)

85. Nguyen, T.D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., Sadeghi, A.R.: Dïot: a federated self-learning anomaly detection system for IoT. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), pp. 756–767. IEEE (2019)
86. Onwuzurike, L., Mariconti, E., Andriotis, P., Cristofaro, E.D., Ross, G., Stringhini, G.: MaMaDroid: detecting Android malware by building Markov chains of behavioral models (extended version). ACM Trans. Privacy Secur. (TOPS) **22**(2), 1–34 (2019)
87. Pajouh, H.H., Dehghantanha, A., Khayami, R., Choo, K.K.R.: Intelligent OS X malware threat detection with code inspection. J. Comput. Virol. Hacking Tech. **14**(3), 213–223 (2018)
88. Parmisano, A., Garcia, S., Erquiaga, M.J.: Stratosphere laboratory. A labeled dataset with malicious and benign IoT network traffic (2020). https://www.stratosphereips.org/datasets-iot23
89. Pascanu, R., Stokes, J.W., Sanossian, H., Marinescu, M., Thomas, A.: Malware classification with recurrent networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1916–1920. IEEE (2015)
90. Peiravian, N., Zhu, X.: Machine learning for Android malware detection using permission and API calls. In: 2013 IEEE 25th International Conference on Tools with Artificial Intelligence, pp. 300–305. IEEE (2013)
91. Pellegrino, G., Lin, Q., Hammerschmidt, C., Verwer, S.: Learning behavioral fingerprints from netflows using timed automata. In: IFIP, pp. 308–316. IEEE (2017)
92. Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., Cavallaro, L.: TESSERACT: eliminating experimental bias in malware classification across space and time. In: 28th USENIX Security Symposium (USENIX Security 2019), pp. 729–746 (2019)
93. Perdisci, R., Lee, W., Feamster, N.: Behavioral clustering of http-based malware and signature generation using malicious network traces. In: NSDI, vol. 10 (2010)
94. Pirscoveanu, R.S., Hansen, S.S., Larsen, T.M., Stevanovic, M., Pedersen, J.M., Czech, A.: Analysis of malware behavior: type classification using machine learning. In: 2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp. 1–7. IEEE (2015)
95. "Polo" Chau, D.H., Wright, A., Nachenberg, C., Faloutsos, C., Wilhelm, J.: Polonium: tera-scale graph mining and inference for malware detection. In: Proceedings of the SIAM International Conference on Data Mining, pp. 131–142. Society for Industrial and Applied Mathematics (2011)
96. Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., Nicholas, C.K.: Malware detection by eating a whole exe. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
97. Rafique, M.Z., Caballero, J.: FIRMA: malware clustering and network signature generation with mixed network behaviors. In: Stolfo, S.J., Stavrou, A., Wright, C.V. (eds.) RAID 2013. LNCS, vol. 8145, pp. 144–163. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41284-4_8
98. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
99. Rieck, K., Trinius, P., Willems, C., Holz, T.: Automatic analysis of malware behavior using machine learning. J. Comput. Secur. **19**(4), 639–668 (2011)

100. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1619–1630 (2006)
101. Ronen, R., Radu, M., Feuerstein, C., Yom-Tov, E., Ahmadi, M.: Microsoft malware classification challenge. arXiv preprint arXiv:1802.10135 (2018)
102. Rosenberg, I., Sicard, G., David, E.O.: DeepAPT: nation-state APT attribution using end-to-end deep neural networks. In: Lintas, A., Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) ICANN 2017. LNCS, vol. 10614, pp. 91–99. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68612-7_11
103. Rosenblum, N., Zhu, X., Miller, B.P.: Who wrote this code? Identifying the authors of program binaries. In: Atluri, V., Diaz, C. (eds.) ESORICS 2011. LNCS, vol. 6879, pp. 172–189. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23822-2_10
104. Sahs, J., Khan, L.: A machine learning approach to Android malware detection. In: 2012 European Intelligence and Security Informatics Conference, pp. 141–147. IEEE (2012)
105. Sami, A., Yadegari, B., Rahimi, H., Peiravian, N., Hashemi, S., Hamze, A.: Malware detection based on mining API calls. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1020–1025 (2010)
106. Santos, I., Brezo, F., Ugarte-Pedrero, X., Bringas, P.G.: Opcode sequences as representation of executables for data-mining-based unknown malware detection. Inf. Sci. **231**, 64–82 (2013)
107. Saxe, J., Berlin, K.: Deep neural network based malware detection using two dimensional binary program features. In: 2015 10th International Conference on Malicious and Unwanted Software (MALWARE), pp. 11–20. IEEE (2015)
108. Saxe, J., Sanders, H.: Malware Data Science: Attack Detection and Attribution. No Starch Press (2018)
109. Sayadi, H., et al.: 2SMaRT: a two-stage machine learning-based approach for run-time specialized hardware-assisted malware detection. In: 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 728–733. IEEE (2019)
110. Sayadi, H., Patel, N., PD, S.M., Sasan, A., Rafatirad, S., Homayoun, H.: Ensemble learning for effective run-time hardware-based malware detection: a comprehensive analysis and classification. In: 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC), pp. 1–6. IEEE (2018)
111. Shabtai, A., Fledel, Y., Elovici, Y.: Automated static code analysis for classifying Android applications using machine learning. In: 2010 International Conference on Computational Intelligence and Security, pp. 329–333. IEEE (2010)
112. Shibahara, T., Yagi, T., Akiyama, M., Chiba, D., Yada, T.: Efficient dynamic malware analysis based on network behavior using deep learning. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–7. IEEE (2016)
113. Simko, L., Zettlemoyer, L., Kohno, T.: Recognizing and imitating programmer style: adversaries in program authorship attribution. Proc. Priv. Enhancing Technol. **2018**(1), 127–144 (2018)
114. Singh, A., Handa, A., Kumar, N., Shukla, S.K.: Malware classification using image representation. In: Dolev, S., Hendler, D., Lodha, S., Yung, M. (eds.) CSCML 2019. LNCS, vol. 11527, pp. 75–92. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20951-3_6
115. Smith, M.R., et al.: Mind the gap: on bridging the semantic gap between machine learning and malware analysis. In: Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security, pp. 49–60 (2020)

116. Souri, A., Hosseini, R.: A state-of-the-art survey of malware detection approaches using data mining techniques. HCIS **8**(1), 1–22 (2018). https://doi.org/10.1186/s13673-018-0125-x
117. Spreitzenbarth, M., Schreck, T., Echtler, F., Arp, D., Hoffmann, J.: Mobile-sandbox: combining static and dynamic analysis with machine-learning techniques. Int. J. Inf. Secur. **14**(2), 141–153 (2015)
118. Suarez-Tangil, G., Dash, S.K., Ahmadi, M., Kinder, J., Giacinto, G., Cavallaro, L.: DroidSieve: fast and accurate classification of obfuscated Android malware. In: Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, pp. 309–320 (2017)
119. Suarez-Tangil, G., Tapiador, J.E., Peris-Lopez, P., Blasco, J.: Dendroid: a text mining approach to analyzing and classifying code structures in Android malware families. Expert Syst. Appl. **41**(4), 1104–1117 (2014)
120. Tamersoy, A., Roundy, K., Chau, D.H.: Guilt by association: large scale malware detection by mining file-relation graphs. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1524–1533. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2623330.2623342
121. Tegeler, F., Fu, X., Vigna, G., Kruegel, C.: BotFinder: finding bots in network traffic without deep packet inspection. In: CoNEXT, pp. 349–360. ACM (2012)
122. Ucci, D., Aniello, L., Baldoni, R.: Survey of machine learning techniques for malware analysis. Comput. Secur. **81**, 123–147 (2019)
123. Verwer, S., Nadeem, A., Hammerschmidt, C., Bliek, L., Al-Dujaili, A., O'Reilly, U.M.: The robust malware detection challenge and greedy random accelerated multi-bit search. In: Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security, pp. 61–70 (2020)
124. Wang, W., Wang, X., Feng, D., Liu, J., Han, Z., Zhang, X.: Exploring permission-induced risk in Android applications for malicious application detection. IEEE Trans. Inf. Forensics Secur. **9**(11), 1869–1882 (2014)
125. Wang, W., Zhu, M., Zeng, X., Ye, X., Sheng, Y.: Malware traffic classification using convolutional neural network for representation learning. In: 2017 International Conference on Information Networking (ICOIN), pp. 712–717. IEEE (2017)
126. Wang, X., Yang, Y., Zhu, S.: Automated hybrid analysis of Android malware through augmenting fuzzing with forced execution. IEEE Trans. Mob. Comput. **18**(12), 2768–2782 (2018)
127. Wang, Z., Tian, M., Jia, C.: An active and dynamic botnet detection approach to track hidden concept drift. In: Qing, S., Mitchell, C., Chen, L., Liu, D. (eds.) ICICS 2017. LNCS, vol. 10631, pp. 646–660. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89500-0_55
128. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
129. Xu, Z., Ray, S., Subramanyan, P., Malik, S.: Malware detection using machine learning based analysis of virtual memory access patterns. In: Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 169–174. IEEE (2017)
130. Yakura, H., Shinozaki, S., Nishimura, R., Oyama, Y., Sakuma, J.: Malware analysis of imaged binary samples by convolutional neural network with attention mechanism. In: Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy, pp. 127–134 (2018)
131. Yang, Y., Wu, L., Yin, G., Li, L., Zhao, H.: A survey on security and privacy issues in internet-of-things. IEEE Internet Things J. **4**(5), 1250–1258 (2017)

132. Ye, Y., Li, T., Adjeroh, D., Iyengar, S.S.: A survey on malware detection using data mining techniques. ACM Comput. Surv. (CSUR) **50**(3), 1–40 (2017)
133. Yerima, S.Y., Alzaylaee, M.K., Sezer, S.: Machine learning-based dynamic analysis of Android apps with improved code coverage. EURASIP J. Inf. Secur. **2019**(1), 4 (2019)
134. Yerima, S.Y., Sezer, S., McWilliams, G., Muttik, I.: A new Android malware detection approach using Bayesian classification. In: 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), pp. 121–128. IEEE (2013)
135. Yokoyama, A., et al.: SandPrint: fingerprinting malware sandboxes to provide intelligence for sandbox evasion. In: Monrose, F., Dacier, M., Blanc, G., Garcia-Alfaro, J. (eds.) RAID 2016. LNCS, vol. 9854, pp. 165–187. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45719-2_8
136. Yuan, Z., Lu, Y., Xue, Y.: DroidDetector: Android malware characterization and detection using deep learning. Tsinghua Sci. Technol. **21**(1), 114–123 (2016)
137. Yuxin, D., Siyi, Z.: Malware detection based on deep learning algorithm. Neural Comput. Appl. **31**(2), 461–472 (2017). https://doi.org/10.1007/s00521-017-3077-6
138. Zhang, F., Chan, P.P., Biggio, B., Yeung, D.S., Roli, F.: Adversarial feature selection against evasion attacks. IEEE Trans. Cybern. **46**(3), 766–777 (2015)
139. Zhang, H., Sun, M., Yao, D., North, C.: Visualizing traffic causality for analyzing network anomalies. In: Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics, pp. 37–42 (2015)
140. Zhang, H., Yao, D.D., Ramakrishnan, N., Zhang, Z.: Causality reasoning about network events for detecting stealthy malware activities. Comput. Secur. **58**, 180–198 (2016)
141. Zhang, H., Zhang, W., Lv, Z., Sangaiah, A.K., Huang, T., Chilamkurti, N.: MALDC: a depth detection method for malware based on behavior chains. World Wide Web **23**, 991–1010 (2019)
142. Zhang, X., et al.: Enhancing state-of-the-art classifiers with API semantics to detect evolved Android malware. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 757–770 (2020)
143. Zhou, Y., Jiang, X.: Dissecting Android malware: characterization and evolution. In: 2012 IEEE Symposium on Security and Privacy, pp. 95–109. IEEE (2012)
144. Zhu, H.J., You, Z.H., Zhu, Z.X., Shi, W.L., Chen, X., Cheng, L.: DroidDet: effective and robust detection of Android malware using static analysis along with rotation forest model. Neurocomputing **272**, 638–646 (2018)
145. Zhu, Z., Dumitraş, T.: FeatureSmith: automatically engineering features for malware detection by mining the security literature. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 767–778 (2016)