

Received 8 August 2022, accepted 30 August 2022, date of publication 5 September 2022, date of current version 12 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3204171

 SURVEY

Explainable Artificial Intelligence in CyberSecurity: A Survey

NICOLA CAPUANO¹, GIUSEPPE FENZA², (Member, IEEE),
VINCENZO LOIA¹, (Senior Member, IEEE),
AND CLAUDIO STANZIONE³, (Member, IEEE)

¹School of Engineering, University of Basilicata, 85100 Potenza, Italy

²Department of Management and Innovation Systems, University of Salerno, 84084 Fisciano, Italy

³Defence Analysis & Research Institute, Center for Higher Defence Studies, 00165 Rome, Italy

Corresponding author: Vincenzo Loia (loia@unisa.it)

ABSTRACT Nowadays, Artificial Intelligence (AI) is widely applied in every area of human being's daily life. Despite the AI benefits, its application suffers from the opacity of complex internal mechanisms and doesn't satisfy by design the principles of Explainable Artificial Intelligence (XAI). The lack of transparency further exacerbates the problem in the field of CyberSecurity because entrusting crucial decisions to a system that cannot explain itself presents obvious dangers. There are several methods in the literature capable of providing explainability of AI results. Anyway, the application of XAI in CyberSecurity can be a double-edged sword. It substantially improves the CyberSecurity practices but simultaneously leaves the system vulnerable to adversary attacks. Therefore, there is a need to analyze the state-of-the-art of XAI methods in CyberSecurity to provide a clear vision for future research. This study presents an in-depth examination of the application of XAI in CyberSecurity. It considers more than 300 papers to comprehensively analyze the main CyberSecurity application fields, like Intrusion Detection Systems, Malware detection, Phishing and Spam detection, BotNets detection, Fraud detection, Zero-Day vulnerabilities, Digital Forensics and Crypto-Jacking. Specifically, this study focuses on the explainability methods adopted or proposed in these fields, pointing out promising works and new challenges.


INDEX TERMS Artificial intelligence, cybersecurity, explainable artificial intelligence, security paradigm, trust.

I. INTRODUCTION

Context. Artificial Intelligence (AI) is becoming more and more prevalent in our daily lives. To quantify this phenomenon numerically, Grand View Research valued the global AI market size at USD 93.5 billion in 2021 and forecasts a compound annual growth rate (CAGR) of 38.1% from 2022 to 2030.¹ Recently, AI finds widely application in many areas as well as in the CyberSecurity domain.

Likewise, Mordor Intelligence valued the global CyberSecurity market at \$156.24 billion in 2020 with an expectation to be worth \$352.25 billion, with an annual growth rate of

14.5%, by 2026.² These numbers help convey the potential of these two fields together and the need to find the proper cohesion. Even if AI algorithms appear effective in outcomes and predictions, they suffer from opacity, making it difficult to gain insight into their internal working mechanisms. This aspect further exacerbates the problem in a field like CyberSecurity because entrusting important decisions to a system that cannot explain itself presents obvious dangers. On the light of this scenario, Explainable Artificial Intelligence (XAI) suggests a transition toward more interpretable AI to overcome this issue. XAI principles intend to develop strategies that will result in better explainable models while keeping high-performance levels.

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

¹<https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market>

²<https://www.mordorintelligence.com/industry-reports/cyber-security-market>

Problems and Motivations. Identifying gaps in the literature to solve the critical issue of CyberSecurity for future ICT systems is critical. The absence of transparency undermines confidence. Security practitioners may hesitate to trust the systems if they do not understand how crucial decisions are made. However, the application of XAI in CyberSecurity can be a double-edged sword: it can substantially improve CyberSecurity practices but it may also facilitate new attacks on the AI applications since it will also be Explainable to the attacker, which may pose severe security threats [1]. As with all innovations, there are pros and cons, but in this case, it seems that the pros outweigh the cons mitigating the risks of AI adoption in analogy to other application domains, like in the Open Source context. Furthermore, the definition of AI models compliant with XAI principles, or the development of model agnostic XAI frameworks, will allow large-scale AI usage in industrial and human scenarios, increasing the capabilities to timely recognize vulnerabilities.

This study aims to compensate for the lack of investigation in this area by focusing on the proposed techniques and how they achieve explainability in order to design a path of promising and appropriate future research directions, hoping that interested researchers will be able to quickly and effectively grasp the key features of the methods analyzed.

Contribution. This paper collects and analyzes the results of an in-depth survey on XAI in CyberSecurity. It aims to take a step back to get a complete picture of the current state of the art in this field of research, considering XAI applications in several areas of CyberSecurity. This work stands out from other works because it focuses on understanding explainability and on comparing explainable and non-explainable procedures used in the most studied areas of CyberSecurity. One of the main points is to provide a solid foundation for further discussion using the lens of the literature.

The main contributions of this paper are:

- A detailed discussion on the main concepts, objectives, and consequences of enabling Explainability in various CyberSecurity applications.
- An organized overview of existing XAI approaches in CyberSecurity, based on a literature review of over 300 papers (an outlook of surveys on XAI, AI in CyberSecurity, and XAI in CyberSecurity is also included).
- A summary tables of the explainable methods analyzed and the most frequently used datasets for each field of application.
- A discussion on past efforts, current trends and future challenges.

Organization. Table 1 presents acronyms used in the document for clarity to be provided to the reader. The rest of the survey is structured as follows. Section II presents an Explainable Artificial Intelligence overview. Section III explores CyberSecurity Threats Foundations and AI applications. Section IV analyzes related surveys, while Section V discusses XAI works in CyberSecurity. Section VI discuss the findings and finally Section VII concludes this survey.

TABLE 1. List of key acronyms.

AI	Artificial Intelligence
ANN	Artificial Neural Network
BRCG	Boolean Rule Column Generation
CAGR	Compound Annual Growth Rate
CEM	Contrastive Explanation Method
CNN	Convolutional Neural Network
CPS	Cyber-Physical Systems
DARPA	Defense Advanced Research Projects Agency
DFF	Deep Feedforward Networks
DGA	Domain Generation Algorithms
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
EBM	Explainable Boosting Machine
ENISA	European Union Agency for CyberSecurity
GA	Genetic Algorithm
GCN	Graph Convolutional Neural Network
GNN	Graph Neural Network
GRAD-CAM	Gradient-weighted Class Activation Mapping
HCI	Human Computer Interaction
HIDS	Host-based Intrusion Detection System
ICT	Information and Communication Technologies
IDS	Intrusion Detection System
IoT	Internet of Things
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
LogRR	Logistic Rule Regression
LORE	Local Rule-based Explanations
LR	Linear Regression
LSTM	Long Short Term Memory
ML	Machine Learning
MLP	Multi Layer Perceptron
NB	Naive Bayes
NIDS	Network Intrusion Detection System
NIST	National Institute of Standards and Technology
NN	Neural Network
OSINT	Open Source Intelligence
PCA	Principal Component Analysis
RCNN	Region Based Convolutional Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanation
SVM	Support Vector Machine
WNIDS	Wireless Network Intrusion Detection System
XAI	EXplainable Artificial Intelligence
XGB	eXtreme Gradient Boosting

II. BACKGROUND ON EXPLAINABLE ARTIFICIAL INTELLIGENCE

DARPA, the Defense Advanced Research Projects Agency, financed the “Explainable AI (XAI) Program” at the beginning of 2017 [2]. XAI aims to develop more understandable models while maintaining a high degree of learning performance (prediction accuracy); and enable human users to comprehend, adequately trust, and manage the future generation of artificially intelligent partners.

After the launch of the program, the scientific contribution in the Explainable Artificial Intelligence field has grown significantly, as shown in Figure 1.

A. XAI TAXONOMY

Throughout the presented literature, various terms have been adopted, trying to cover all possible fields of application. Following are just a few of the wide variety used:

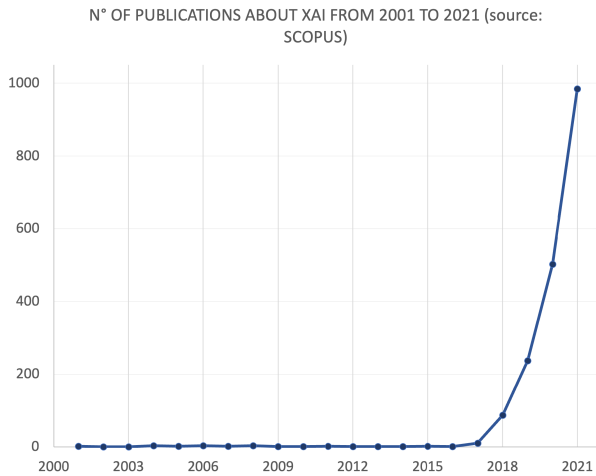


FIGURE 1. Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI until 2021. Data retrieved from Scopus using as search key [TITLE-ABS-KEY (Explainable AND Artificial AND Intelligence)].

Transparency: Do users grasp the format and language choices made by the model?

Fairness: Can it be proven that model judgments are fair to protected groups?

Trust: How comfortable are human users with using the system?

Usability: How well-equipped is the system to give users a secure and productive environment in which to complete their tasks?

Reliability: How resistant is the system to changes in parameters and inputs?

Causality: Do the predicted changes in the output, resulting from input perturbation, occur in the actual system?

In the middle of 2020, the National Institute of Standards and Technology (NIST) presented four fundamental principles for explainable AI systems [3] as shown in Figure 2. The *Explanation* principle obligates AI systems to supply evidence, support, or reasoning for each output. A system fulfils the *Meaningful* principle if the recipient understands the system’s explanations. The *Explanation Accuracy* principle imposes accuracy on a system’s explanations and in the end *Knowledge Limits* principle states that systems identify cases they were not designed or approved to operate, or their answers are not reliable [3].

Over the years, a vast taxonomy has been developed on the various ways and methods that can make an AI model explainable. The first distinction needed is between *Interpretability* and *Explainability*. *Interpretability* is all about understanding the cause and effect within an AI system. On the other hand, *Explainability* goes beyond interpretability in that it helps us understand how and why a model came up with a prediction in a human-readable form. Figure 3 presents the current taxonomy and makes a crucial distinction between true transparency (interpretable models) and post-hoc interpretations (additional techniques used to

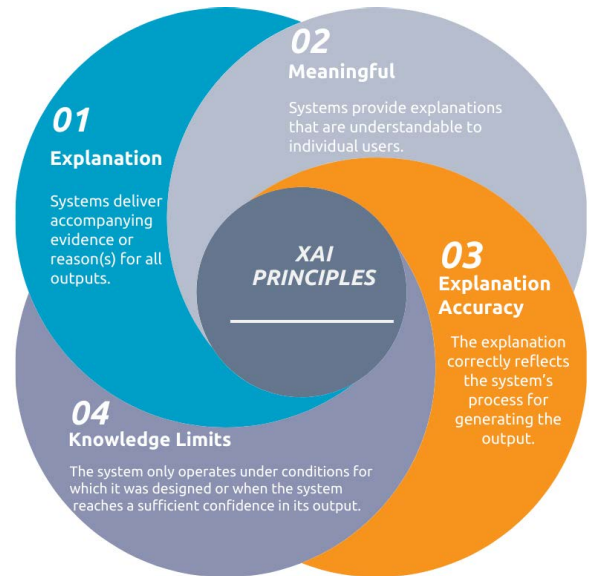


FIGURE 2. XAI Principles presented by NIST in [3].

shed transparency on complex black-box models). These techniques include producing local explanations for specific inputs or the entire model globally. Following a quick overview:

- **Model Specific or Model Agnostic:** This determines whether or not the interpretation method is restricted to a specific model. Model-specific methods and tools are those that are specific to a model. Model agnostic methods can be applied to any ML model to gain interpretability. Internal model data such as weights and structural details are not accessible to these models.
- **Intrinsic or Extrinsic (post-hoc):** This indicates whether the model is interpretable on its own or whether interpretability requires using methods that examine models after training. Simple, comprehensible models, like decision trees, are intrinsic. Utilizing an interpretation strategy after training to achieve interpretability is extrinsic.
- **Local or Global:** Whether the interpretation method describes a single data record or all of a model’s behaviour depends on whether it is local or global. Global methods and tools interpret the entire model, whereas Local methods and tools only explain a single prediction.

B. XAI FRAMEWORKS

An XAI framework is a tool that creates reports on model activity and tries to explain how it works. The following are the main ones encountered during the Survey.

LIME. Local Interpretable Model-agnostic Explanations (LIME) is a framework that seeks to provide an individual-level explanation of individual predictions (*Local*) in an extrinsic (*Post-hoc*) manner and is able to explain any model without needing to ‘peak’ into it

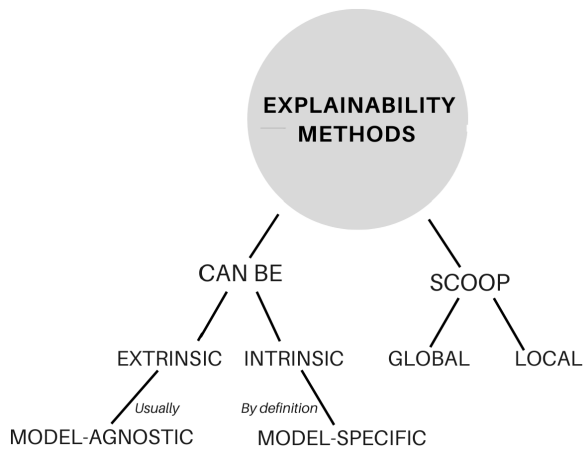


FIGURE 3. A visual representation of XAI taxonomy.

(*Model-Agnostic*) [4]. In order to figure out what parts of the interpretable input are contributing to the prediction, it perturbs the input around its neighbourhood and see how the model's predictions behave. Then it generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model. On this new dataset, LIME then trains an interpretable model, which is weighted by the proximity of the sampled instances to the instance of interest.

SHAP. SHapley Additive exPlanations (SHAP) [5] is a framework with a clear goal, explaining the prediction of an instance x by computing the contribution of each feature to the prediction. Like LIME, it is a *Local-based*, *Post-hoc*, and *Model-Agnostic* paradigm. The SHAP explanation technique uses coalitional game theory to compute Shapley values. A data instance's feature values operate as coalition members. Shapley values inform how fairly distributed the prediction is across the characteristics. A player might be a single feature value or a collection of feature values. It is not necessary to establish a local model in SHAP (as opposed to LIME), but rather the same function is used to calculate the Shapley values for each dimension.

Anchors. The Anchors approach [6] locates a decision rule that “anchors” the prediction adequately and uses it to explain specific predictions of any black box classification model. If changes in other feature values do not affect the prediction, a rule anchors it. Anchors reduces the number of model calls by combining reinforcement learning techniques with a graph search algorithm. The ensuing explanations are expressed as simple IF-THEN rules known as anchors. This framework is *Local-based*, *Post-hoc* and then *Model-Agnostic*.

LORE. Local Rule-based Explanations (LORE) [7] creates an interpretable predictor for a given black box instance. A decision tree is used to train the local interpretable predictor on a dense set of artificial cases. The decision tree allows for the extraction of a local explanation, which consists of a single choice rule and a collection of counterfactual rules for the reversed decision. This framework is *Local-based*, *Post-hoc* and then *Model-Agnostic*.

GRAD-CAM. Gradient-weighted Class Activation Mapping (GRAD-CAM) [8] is a technique for producing a class-specific heat map from a single image. Grad-CAM produces a class discriminative localization map as a result. The framework makes use of the feature maps generated by a CNN's final convolutional layer. This is *Local-based*, *Post-hoc* but *Model-Specific*.

CEM. Contrastive Explanation Method (CEM) [9] provides explanations for classification models. More in detail, it retrieves the features that should be sufficiently present to predict the same class for the input instance. It also identifies minimal features to change for associating the input instance to a different class. This is *Local-based*, *Post-hoc* but *Model-Agnostic*.

III. CYBERSECURITY THREATS FOUNDATIONS AND AI APPLICATIONS

If it were measured as a country, Cybercrime, which inflicted damages around \$6 trillion globally in 2021, would be the world's third-largest economy after the U. S. and China. CyberSecurity Ventures expects global cybercrime costs to grow by 15% per year over the next five years, reaching \$10.5 trillion annually by 2025, up from \$3 trillion in 2015. In addition to being exponentially more considerable than the damage caused by natural disasters in a year, this represents the most significant transfer of economic wealth in history and poses a threat to the incentives for innovation and investment [10].

CyberSecurity is the process of defending ICT systems against various cyber threats or attacks. A “cyber-attack” is any criminal activity that preys on electronic information systems, networks, or infrastructure. Information is primarily intended to be stolen, altered, or destroyed. In the current cyber-attack situation, attack vectors that take advantage of a lack of readiness and (system as well as human) preparedness to access sensitive data or compromise systems are frequent. The main problems of CyberSecurity are the knowledge of various cyber-attacks and the development of complementary protection mechanisms.

The risks usually connected to any attack take into account three security variables: threats, who is attacking; vulnerabilities, or the holes they are attacking; and impacts, or what the assault does. A security incident is an act that threatens the confidentiality, integrity, or availability of information assets and systems. Obtaining illegal access, destruction, and alteration of information to harm possibly are just a few examples of potential breaches and security violations on a computer system or mobile device. Threats describe all of the security mentioned above infractions' potential risk and hazard, and attacks describe any attempts to commit a violation.

Measures to safeguard information and communication technology, the unprocessed data and information it contains, as well as their processing and transmission, associated virtual and physical elements of the systems, the degree of protection attained as a result of the application of those

measures, and ultimately the associated field of professional endeavour, are all associated with CyberSecurity.

Cyber-attacks or intrusions require defence techniques to protect data or information, information systems, and networks. They are in charge of preventing data breaches and security incidents, as well as monitoring and responding to intrusions, defined as any unauthorized action that causes damage to an information system.

ENISA, the European Union Agency for CyberSecurity, provided a report with an analysis of the top 15 cyber threats, showed in Figure 4, that dominated the period between January 2019 and April 2020 [11].

Only some of these threats were addressed in this survey, focusing on those application areas where Explainable Artificial Intelligence has been most explored. In particular, the world of Intrusion Detection Systems, Malware detectors, prevention against Spam and Phishing, and detection of BotNets was extensively explored. In addition, a shorter analysis was conducted on *Fraud Detection, Zero-Day Vulnerabilities, Digital Forensics, Cyber-Physical Systems* and *Crypto-Jacking*.

IV. RELATED WORKS

The following sub-sections analyze the existing surveys related to this work. First, there is an analysis of existing surveys in the general field of Explainable Artificial Intelligence. Subsequently, attention will be focused on surveys about AI applications in CyberSecurity. To conclude, there is an investigation of the few existing works that attempt to clarify the applications of Explainable Artificial Intelligence in CyberSecurity.

A. SURVEYS ON EXPLAINABLE ARTIFICIAL INTELLIGENCE

High-performance AI systems, particularly those based on DL, behave similarly to black boxes that provide good results but can hardly justify a given output in a human-understandable way [12], [13]. It is essential to minimize potential biases (e.g., algorithmic, racial, ideological and gender biases) during the ethical AI solution development stage [14], [15].

Adadi and Berrada [16] conducted an exhaustive literature analysis, collecting and analyzing 381 different scientific papers between 2004 and 2018. They organized all of the scientific work in explainable AI along four primary axes and emphasized the importance of introducing more formalism in the field of XAI and more interaction between people and machines.

Abdul et al. [17] evaluated a large corpus of explainable research based on 289 core papers and 12412 citing publications and created a citation network to set an HCI (Human Computer Interaction) research agenda in Explainability. This work focused primarily on developing an HCI research agenda in Explainability and investigating how HCI research might aid in the development of existing explainable systems that are effective for end-users. Staying on the subject of visualization for XAI, [18] provides a comprehensive assessment

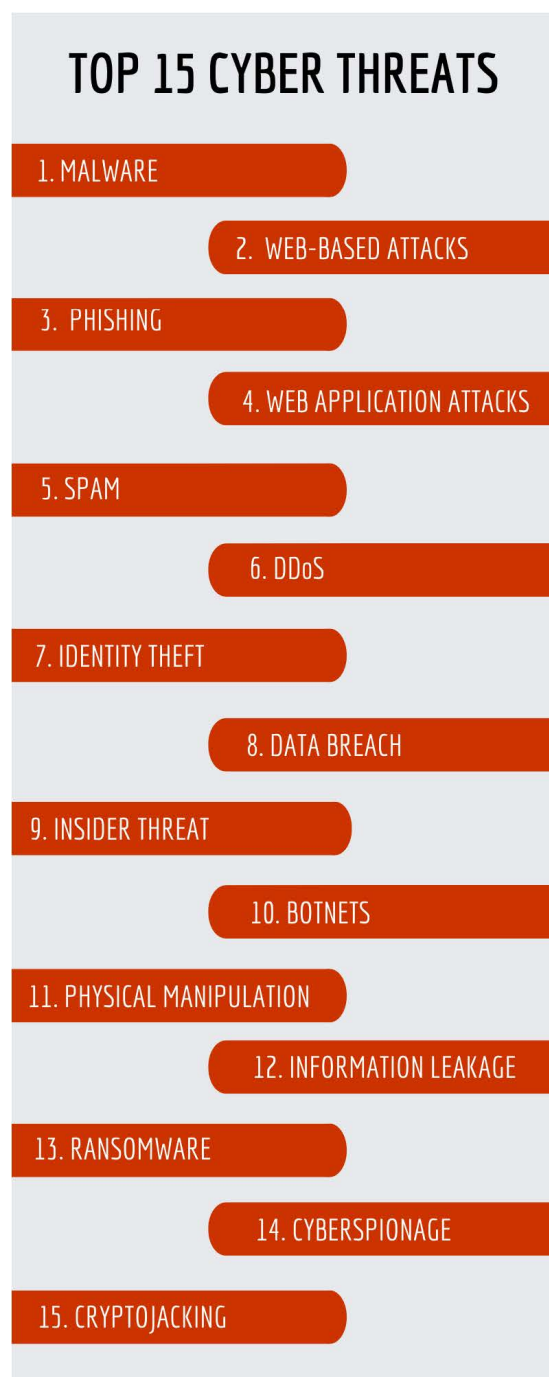


FIGURE 4. Top 15 Cyber Threats presented by ENISA in [11].

of recent studies on visual interpretability of neural networks, covering visualization and diagnosis of CNN (Convolutional Neural Network) representations, techniques for disentangling CNN representations into graphs or trees, and learning of CNNs with disentangled and interpretable representations ending with a middle-to-end learning based on model interpretability.

The authors of [19] employed a loss for each filter in high-level convolutional layers to force each filter to learn

extremely particular object components to improve the interpretability of traditional CNNs. Also, Angelov *et al.* [20] cover the visualization technique; in particular, they proposed a broader taxonomy, considering whether the explanation is local or not, if the models are transparent or opaque, if the techniques are model-specific or model-agnostic, and whether explanations are created by simplification, conveyed through visualizations or based on feature relevance. In the same line, one of the works worth mentioning is that edited by Arrieta *et al.* [21], which developed a new style of organization that first distinguishes between transparent and post-hoc approaches and then creates sub-categories.

A methodological approach for evaluating the interpretability of ML models is proposed in [22], based on a taxonomy that separates three forms of Explainability: imitate the processing, explain the representation, and explain-producing networks. Methods for describing black-box models on a wide scale, such as data mining and ML, were reviewed in [23]. They provided a full taxonomy of Explainability strategies based on the problem they were dealing with.

In [24] are examined and presented several XAI approaches, validation measures, and the types of explanations that can be generated to improve the acceptance of expert systems among general users.

The authors in [25] focus on machine interpretation in the medical industry and reveal the difficulty of reading a black box model's choice.

In philosophy and sociology, Mittelstadt *et al.* [26] pay attention to the differences between these models and explanations.

Miller's work [27] is likely the most important attempt to articulate the connection between human science and XAI. Miller gave an in-depth assessment of studies on the explanation problem in philosophy, psychology, and cognitive science in his paper. According to the author, the latter could be a vital resource for the advancement of the field of XAI.

In [28], the attention is focused on the fidelity of work closely related to the explanation accuracy. The authors surveyed several studies that have evaluated explanation fidelity.

Predictive accuracy, descriptive accuracy, and relevancy are three types of metrics presented by the Predictive, Descriptive, and Relevant (PDR) framework for evaluating interpretability methodologies [29]. They discussed transparent models and post-hoc interpretation, believing that post-hoc interpretability could improve a model's predictive accuracy and that transparent models could expand their use cases by increasing predictive accuracy, demonstrating that the combination of the two methods is ideal in some cases.

As presented in [30], an alternative perspective on hybrid XAI models entails augmenting black-box model expertise with that of transparent model.

The stages are ante-hoc and post-hoc, according to Vilone and Longo [31], [32]. In general, ante-hoc methods consider generating the rationale for the decision from the very beginning of the data training to achieve optimal performance.

An external or surrogate model and the base model are used in post hoc approaches. The base model remains unmodified, while the external model generates an explanation for the users by mimicking the behavior of the base model. In addition, post hoc approaches are classified into two groups: model-agnostic and model-specific. Model-agnostic methods can be used with any AI/ML model, but model-specific approaches only apply to certain models.

Carvalho *et al.* [33] add a criterion on the stage of model development, in-model interpretability that concerns ML models that have inherent interpretability in it (through constraints or not). The need to consider the perspectives of diverse stakeholders is highlighted in [34]. As a result, explanations should be adapted to the particular audience for which they are intended to deliver the relevant information. In [35] a survey of XAI methods in deployment is made, and [36] which considers the XAI for tabular data. To end this review of works in Explainable Artificial Intelligence it is worth considering also [37] where are identified future research directions with Explainability as the starting component of any AI system.

In this section, only works published in the last 5 years, i.e., from 2018 to 2022, have been analysed. However, these works are focused only on the survey of XAI methods emphasizing the most common ones and the general requirements of explainability that are different in CyberSecurity context.

B. SURVEYS ON ARTIFICIAL INTELLIGENCE APPLICATIONS IN CYBERSECURITY

This section presents works that survey the existing literature on AI applications in the world of CyberSecurity. AI and ML play a substantial role in the protection of computer systems [13], [38], [39], [40], [41].

The interaction of AI and CyberSecurity was discussed by the author in [42]. The study looked, in particular, at ML, and DL approaches to countering Cyber threats [43].

There are various advantages and disadvantages to the use of AI in this field, as briefly analyzed in [44] and [45], and work like that done in [46], where all the existing literature on the last decade is analyzed, can be of help to those who are entering into the specific sector.

Sarker *et al.* [47] proposed a broad definition of CyberSecurity that takes into account all relevant definitions. Information Security, Network security, operational security, application security, Internet of Things (IoT) Security, Cloud security, and infrastructure Security are all covered by CyberSecurity [48].

In [46], more than 770 papers were analyzed, and an overview of the challenges that ML techniques face in protecting Cyberspace against attacks was provided by presenting literature on ML techniques for CyberSecurity, including intrusion detection, spam detection, and malware detection on computer and mobile networks.

Related to this, Gupta *et al.* [49] provide a thorough examination of the various ML and DL models used in mobile network electronic information Security.

The main distinction that came up when analyzing the literature on this subject is the use of ML or DL techniques. In [50] and [51], both cases are analyzed with an in-depth analysis of the various techniques used. Furthermore, both papers specify that only the last three years of literature have been considered, showing that it is a field that has been receiving attention for not very long.

Shaukat *et al.* [52] examined the performance of various ML algorithms in terms of time complexity for identifying Cyber-attacks. The authors focused on fraud detection, intrusion detection, spam detection, and virus detection during their investigation.

Alabadi and Celik in [53] presented a comprehensive survey about using CNN as a key solution for anomaly detection.

Kim and Park [54] focus the attention on ML in Cyber-Physical Systems (CPS), which is the integration of a physical system into the real world and control applications in a computing system, interacting through a communications network. They suggest a CPS structure that divides the system's functions into three layers: physical, network, and software applications. In the sphere of CyberSecurity, researchers apply DL techniques for a variety of applications such as detecting network intrusions, malware traffic detection and classification, and so on, as analyzed extensively in [55], [56], [57], and [58].

The performance of seven DL models on the CSE-CIC-IDS2018 and Bot-IoT datasets is examined in [59]. The models are evaluated on two datasets in this benchmark, and three evaluation metrics are reported. The whole execution of the study is made public in order to facilitate objective comparisons and transparency in [60]. For the specific field of phishing interesting approach is defined in [61] and for ransomware attacks in [62].

Also in this section, only works published in the last 5 years, i.e., from 2018 to 2022, have been analysed. However, these works are focused only on the survey of CyberSecurity threats and methods.

C. XAI SURVEYS IN CYBERSECURITY

Compared to the previous two sections, few works focus on and survey XAI methods in CyberSecurity. Currently, only two work focus exclusively on this area, which are [63], [64]. However, it must be pointed out that in [63], the authors provide a quick overview and, above all, do not pay attention on the different applications within CyberSecurity. In [64] the authors focus on application of XAI in CyberSecurity for specific vertical industry sectors, namely in smart healthcare, smart banking, smart agriculture, smart cities, smart governance, etc..

Exciting work is [65] where the authors made three contributions: a proposal and discussion of desiderata for the explanation of outputs generated by AI-based CyberSecurity systems; a comparative analysis of approaches in the literature on Explainable Artificial Intelligence (XAI), and a general architecture that can serve as a roadmap for guiding research efforts towards AI-based CyberSecurity systems.

In [66] Vigano *et al.* presented Explainable Security (XSec), a new security paradigm that involves several different stakeholders and is multifaceted by nature. In [67] the authors carried out a comprehensive literature review of various DL architectures applied in CyberSecurity, including state-of-the-art studies conducted with explainable AI. Indeed, [68] focuses on Android Malware Defenses and XAI applications in this field; they point out that nine out of ten primary sources are proposed after 2019, indicating that Explainable Deep Learning approaches for malware defenses are a current hot research topic.

Works analysed in this section are in the last 3 years, i.e., from 2020 to 2022. Although all of these publications are outstanding, none demonstrate how explainability occurs in key sectors of AI in CyberSecurity, which is the primary focus of this survey.

V. LITERATURE REVIEW

In the following subsections, the works that seek to achieve explainability in the field of CyberSecurity were reviewed. In particular, the discussion focuses on the following application fields:

- Intrusion Detection Systems
- Malware Detection
- Phishing and Spam Detection
- BotNet Detection

The template used for describing the results of the analysis of the works falling in the above application fields is this:

- *Brief Introduction*, a small analysis of the specific topic;
- *Why XAI*, a motivation based mostly on data, for why Explainable Artificial Intelligence is needed in that particular domain;
- *State of art of AI methods*, a quick look at applied AI methods;
- *State of the art of Explainable Artificial Intelligence*, an exhaustive analysis of existing XAI methods with a specific focus on the explainability method;
- *Consideration*, a brief discussion of the analysis carried out and an overview of the main directions explainable methods are moving.

In addition to the CyberSecurity applications aforementioned above, other fields will be treated with lesser level of detail, due to the availability of a fewer number of works, focusing only on the review of works using XAI, that are: *Fraud Detection*, *Zero-Day Vulnerabilities*, *Digital Forensics*, and *Crypto-Jacking*.

All application fields were selected according to the relevance and volume of literature to the current state of the art.

A. INTRUSION DETECTION SYSTEMS

Intrusion Detection Systems enable continuous security monitoring of a cyber perimeter in order to timely identify attacks on computers and computer networks.

IDSs can be implemented with hardware appliances or with special software; sometimes, they combine both systems [69].

They do not replace firewalls but integrate them to provide more comprehensive protection. The purpose of the firewall is to selectively (and “mechanically”) intercept data packets (according to a set of predefined rules that packets must follow in order to enter or leave the local network). Traditional firewalls operate on the lowest layers of network communication, thus with filtering rules limited to IP addresses, ports, time of day and a few other criteria [70].

IDSs, on the other hand, are placed “downstream” of the firewall and analyze data packets and the behaviour they generate. Therefore, if an attack originates within the local network, the firewall will not be able to block it. At the same time, the IDS can detect anomalous situations.

IDS systems can be divided into two categories depending on where the intrusion-detection sensors are placed (on the network or a host/endpoint).

Network-based IDS systems (NIDS) analyze IP packets, policing the entire network data traffic. This way, they can complement the firewall where it does not block packets due to misconfiguration or unrestrictive rules; they can also monitor the behaviour of users inside the network.

Host-based intrusion detection systems (HIDS) are typically tools that are installed on a machine (host) and are intended to protect a specific PC (a kind of “super-antivirus”). They can also integrate firewall functions, sandboxing, and so on.

Another distinction can be made in detecting and alerting approaches, which are *Signature-based* and *Anomaly-based*. While *Signature-based* detection is used to detect known threats, *Anomaly-based* detection detects changes in behaviour. *Signature-based* detection is based on a predefined set of known Indicators Of Compromise (IOCs). Malicious network attack behaviour, email subject line content, file hashes, known byte sequences, or malicious domains are all examples of IOCs. Signatures may also include network traffic alerts, such as known malicious IP addresses attempting to access a system. Unlike *Signature-based* detection, *Anomaly-based* detection can discover unknown suspicious behaviour. Anomaly detection begins by training the system with a normalized baseline and comparing activity to that baseline.

1) WHY XAI IN IDSs?

In BakerHostetler’s 2021 Data Security Incident Report,³ some interesting numbers help to understand why the collaboration of AI and humans is needed to combat an already huge problem. 58 % of detected incidents are attributable to Network Intrusion, the most significant cause among the top 5.

On average, in 2020 were needed 92 days to discover the presence of an intrusion, 6 days to contain it, 42 days for forensic efforts to complete, and 90 days total from the date of discovery to notification to end-user. Figure 5 shows the

³<https://www.bakerlaw.com/webfiles/Privacy/2021/Alerts/2021-DSIR-Report.pdf>

N° OF PUBLICATIONS ABOUT IDS FROM 2000 TO 2021 (source: SCOPUS)

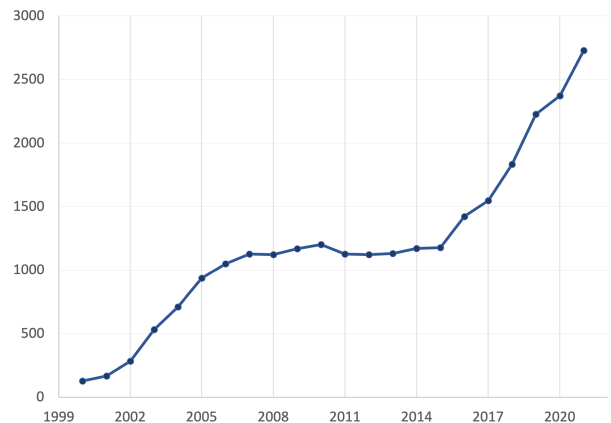


FIGURE 5. IDS Publications from 2000 to 2021, retrieved from Scopus using as search key [TITLE-ABS-KEY (intrusion AND detection AND systems)].

increasing trend of publications in this area. Most of these developed in recent years are based on Machine and Deep Learning algorithms.

The approach based on ML and DL automates the analytical process to find intrusions. High performance, adaptability, flexibility, and the capacity to identify zero-day assaults are the significant benefits of the ML technique. However, there are some drawbacks to ML-based IDS, including high bias propensity, inability to manage outliers, difficulties handling huge datasets, and complex data preprocessing.

The DL-based approach can handle dynamic data that changes over time, recognize large-scale and multi-dimensional data and identify anomalies in the data. Nevertheless, DL-based approaches have many drawbacks, such as a lack of flow information, vulnerability to evasion attempts, poor data knowledge required to design relevant features, and a lack of qualified domain experts to review the implementation. These very latter two points lead back to the need for explainability, a need shared by any agent attempting to give an explanation for the model result and be able to improve it consequently.

2) ARTIFICIAL INTELLIGENCE IN IDSs

Chawla et al. [71] propose a *Host-based* IDS that uses sequences of system calls to identify the expected behaviour of a system. The work describes an efficient *Anomaly-based* intrusion detection system based on CNN layers to capture local correlations of structures in the sequences and Gated Recurrent Units layer to learn sequential correlations from the higher level features.

By examining Linux kernel 5.7.0-rc1, the authors of [72] bridge the gap between theoretical models and application settings. This environment investigates the viability of *HIDS* in modern operating systems and the constraints placed on *HIDS* developers. Keeping the focus on *HIDS* in [73], Gassais et al. propose a framework for intrusion detection in

IoT which combines user and kernel space using AI techniques to automatically get devices behavior, process the data into numeric arrays to train several ML algorithms, and raise alerts whenever an intrusion is found. In [74] and [75] the authors focus the attention on Cloud Environment by detecting *Anomalies* while [76] propose a Siamese-CNN to determine the attack type converting it to an image.

Analyzing the *Network-based* approaches, in [77], the authors present a *NIDS* model that employs a non-symmetric deep AutoEncoder and a Random Forest classifier. Using a non-symmetric deep Auto Encoder for efficient feature selection reduces the model's complexity, similar to [78] and [79] where the classifier is the Support Vector Machine.

Ali et al. in [80] use a Fast Learning Network with a Swarm optimization algorithm, similar to the works in [81] and [82]. The most recent work brings the spotlight on the use of Neural Networks [83], [84] and Adversarial Methods [85], [86], [87].

3) EXPLAINABLE ARTIFICIAL INTELLIGENCE IN IDSs

In [88], a system is proposed that is based on rules dictated by experts. It is *Hybrid* in the sense that it is a combination of human work and ML. The Explainability comes from *Rule-based*; the model behind it is a Decision Tree, a white-box model.

Szczepanski et al. in [89] propose a combination of oracle (ML model, in this case, tested ANN with a PCA) and an explainer module that would explain why a given classification is made. In the explainer module, one compares the distance from the clusters created on the training data. Then, the cluster closest to the test set instance is used for explanation.

In [90], the idea is to use an adversarial approach in order to be able to account for the minimal changes necessary for a classifier to arrive at an incorrect classification. The method thus makes it possible to visualize the features responsible for misclassification. For example, regular connections with low duration and low login success are misclassified as attacks. In contrast, attack connections with a low error rate and higher login success are misclassified as regular, demonstrating that relevant features significantly affect the final result.

A new way of interpreting an Intrusion Detection System is presented in [91]. The authors propose the use of SHAP for both local and global explanations. SHAP, by its nature, is a local method; they propose combining all local explanations to obtain a global explanation of the model. Almost equal work, with some less experimentation, is proposed in [92]. Le et al. [93] propose similar work through SHAP with an ensemble Tree model given a Decision Tree and a Random Forest model. Specifically, at the global level, they use a Heatmap for visualizing the impact of individual features on the classification of the overall model. At the local level, they use a Decision Plot to explain decisions on individual instances of the datasets. Another similar work is the framework proposed by [94], consisting of a Random Forest model using SHAP. The model can assess the credibility of the predicted results and ensure a high level of accuracy in detecting

modern Cyber threats. The strategy adopted makes the final decision after cross-validation of the local explanation of the predicted outcome with the global explanation of SHAP.

The general idea proposed in [95] against adversarial attacks is divided into two parts, initialization and detection. During initialization, the model is trained with an SVM and features and characteristics that make a Normal classification are deduced via LIME. During detection, the Intrusion Detection System goes to compare. If it does not find the data as Normal, it classifies as an attack. On the other hand, if it is classified as Normal, there is a risk of an adversarial attack that is fooling the model. So a further check is done by reusing LIME. After that, the final result is reached.

FAIXID [96] is a new proposed framework that uses data cleaning techniques. They used four algorithms in the experiment to make the results explainable. They use the Boolean Rule Column Generation (BRCG) algorithm [97], which provides a directly interpretable supervised learning method for binary classification. Logistic Rule Regression (LogRR) [98] is a directly interpretable supervised learning method that can perform logistic regression on rule-based functions. The ProtoDash algorithm [99] provides example-based explanations to summarize datasets and explain the predictions of an AI model. Finally, the Contrastive Explanations Method (CEM) is used to compute explanations that highlight both relevant positives (PP) and relevant negatives (NP). Their proposal is not static but involves the use of algorithms depending on the specific case.

The work proposed in [100] defines a method to make rules for accessing the network dynamically and not statically as, for example, the rules set in a firewall may be. Thus, Explainability is the focus of the proposal. The explanation of the results consists of two main steps: i) training a model to approximate the local decision boundary of the target predictive model, and ii) reasoning about the trained model and the given input based on an explanation logic. The explanation is Local-based. They are inspired by LEMNA [101].

The aim in [102] is to increase transparency in an IDS based on a Deep Neural Network. Feedback is presented by computing the input features most relevant to the predictions made by the system. The model adopted is an MLP. Two forms of feedback are generated: 1) offline feedback (after training, before deployment) and 2) online feedback (during deployment). In offline feedback, the user is given the most relevant input features for each concept learned from the system. This information allows the user to evaluate whether the input characteristics that guide the IDS's decision toward a particular class (i.e., the type of attack) align with the domain experts' knowledge. On the other hand, the user is given the most relevant input characteristics for each prediction in the online feedback.

In [103], the authors focus on the possibilities of analyzing encrypted traffic, particularly for accurate detection of DoH (DNS Over HTTPS) attacks. They implement an explainable AI through the use of SHAP that allows visualizing the contribution of individual features to the model classification

decision. Similarly, EXPLAIN-IT [104] is applied to the YouTube video quality classification problem in encrypted traffic scenarios. The work is based on a methodology that deals with unlabeled data, create meaningful clusters and proposes an explanation of the clustering results to the end-user. They use LIME interpreting clusters that are associated with a Local-based strategy then. Alike, ROULETTE [105] focuses on Network traffic. Specifically, attention is coupled with a multi-output DL strategy that helps better discriminate between network intrusions categories. As Post-hoc explanations, they consider visual explanation maps produced through Grad-CAM.

A two-stage ML-based Wireless Network IDS (WNIDS) is implemented in [106] to improve the detection of impersonation and injection attacks in a Wi-Fi network. The XAI was implemented to gain insight into the decisions made by the first-stage ML model, especially for cases where records were predicted as impersonation or injection. The features that contribute significantly to their prediction were determined. This set of features almost corresponds to those identified by the feature selection method for the second-stage ML model. They use SHAP.

In [107], the authors create a framework with a Deep Neural Network at its base and apply an XAI method depending on who benefits from it. For data scientists, SHAP and BRCG [97] are proposed, while for analysts Protodash is used. For end-users where an explanation on the single instance is required, they suggest SHAP, LIME, and CEM. Saran *et al.* [108] propose a comparison between the NetFlow-based feature set⁴ and the feature set designed by the CICFlowMeter tool.⁵ This reliable comparison demonstrates the importance and need for standard feature sets among NIDS datasets, such as evaluating the generalizability of ML model performance in different network environments and attack scenarios. The SHAP method is used to explain the prediction results of ML models by measuring the importance of features. For each dataset, key features that influence model predictions were identified.

In conclusion, this work mentions [109], where an explainable automotive intrusion detection system is proposed, and [110] where a new general method is presented and tested on an IDS dataset. In [111] instead, the authors emphasize the importance of trust but do not use XAI methods.

4) CONSIDERATIONS ABOUT IDS AND XAI

It is interesting to note that most of the methods analyzed use already developed methods to make the results explainable, so the explanation is post-hoc. In particular, in the case of methods already in the research landscape, SHAP is the most adopted method. LIME, on the other hand, has been adopted in only one case. Some frameworks are white-box in nature; most are based on a decision tree.

It would be good to consider frameworks with intrinsic interpretability and not the application of methods for a post-hoc explanation. Furthermore, the final output should be aimed at precise figures and not just any user, such as analysts and defenders. To be explored for future research is the topic of adversarial attacks where the collaboration between humans and machines is necessary and explanations are fundamental to combat this type of intrusion.

B. MALWARE DETECTION

The term malware refers to programs potentially harmful to the user, which are aimed at stealing sensitive data, controlling the PC, or stealing user identity. The term malware originates from the contraction of the words “malicious software” and stands for a program (an executable, a dynamic library, a script, an HTML page, a document with macros, etc.) having unwanted and potentially dangerous effects on the user such as stealing sensitive data, controlling activity at the PC, identity theft, encrypting the hard disk with subsequent ransom demands, and so on.

Malware is usually classified according to its behaviour as Botnet, Backdoor, Information Stealer, Downloaders, Scareware, Rootkit, Worm, Virus, Ransomware or Trojans.

Some of the most common methods an attacker uses are Spam, Phishing, Hacking, Banner advertising, Search page rank, Expired domains or Domain Name Server (DNS) hijacking.

Malware detection techniques can be classified into three main categories (although other classifications exist): (i) *Signature-based*, (ii) *Anomaly-based*, and (iii) *Heuristic-based*.

When using a *Signature-based* approach, programmers scan a file for malware, compare the information with a database of virus signatures, and then verify the results. If the information matches the information in the database, the file is infected with viruses. This approach limits the detection of unknown malware, but its main advantage is that it works well for known malware.

Anomaly-based methods mitigate the limitations of signature-based techniques, allowing detection of any known or unknown malware by applying classification techniques to the actions of a system for malware detection. Detection of malware activity is improved by moving from pattern-based to classification-based detection to identify normal or anomalous behaviour. Applying AI to *Signature-based* and *Anomaly-based* detection systems improves the efficiency of malware detection. *Heuristic-based* method use data mining and ML techniques to learn the behavior of an executable file.

1) WHY XAI IN MALWARE DETECTION?

According to AV-Test Institute,⁶ more than 1 billion malware programs are out there, and 560, 000 new pieces of malware are detected every day. Statista detected that 68.5% of

⁴<https://en.wikipedia.org/wiki/NetFlow>

⁵<https://github.com/CanadianInstituteForCyberSecurity/CICFlowMeter>

⁶<https://www.av-test.org/en/statistics/malware/>

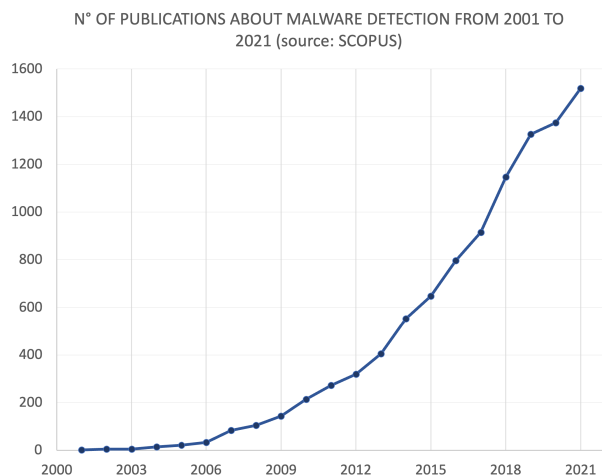


FIGURE 6. Malware Detection Publications from 2001 to 2021, retrieved from Scopus using as search key [TITLE-ABS-KEY (malware AND detection)].

businesses were victimized by ransomware in 2021, a considerable increase from the previous three years. Overall, the number of detected malware types stood at 28.84 million in 2010; by 2020, this had reached nearly 678 million.⁷ Figure 6 shows the increasing trend of publications in this area, reflecting its considerable attention. XAI can assist with risk identification and prioritization, incident response coordination, and malware threat detection. XAI appears to be a good answer in situations demanding explainability, interpretability, and accountability, where humans require assistance in fighting a massive number of attacks.

2) ARTIFICIAL INTELLIGENCE IN MALWARE DETECTION

In [112], the authors propose an *Anomaly-based* approach where the system employs significant features of activity to model normal and malicious behaviour of users in Cloud-based environments. Similar are the works in [113] and [114] where extreme surveillance through malware hunting is delivered. Keeping with *Anomaly-based* approaches, Alaeiyan et al. introduce [115] VECG, a tool for exploring and supplying required environmental conditions at runtime, while in [116] Stiborek et al. propose a novel tool that detects malware observing the interactions between the operating systems and network resources.

ASSCA [117] is a system architecture that combines the DL model based on sequence data and the ML model based on API statistical features, similar to what happens in [118] where the API call relation is extracted, the ordered cycle graph is constructed based on Markov chain and then the graph convolution neural network (GCN) detects malware. Other exciting works based on DL of Behavior Graphs are [119], [120] where for the detection are used file content and file relations.

⁷<https://www.statista.com/topics/8338/malware/dossierKeyfigures>

Staying on the use of graphs but moving to *Signature-based* systems, HLES-MMI [121] is a method that identifies metamorphic malware families based on computing the similarities among the higher-level engine signatures. Khan et al. [122] analyzed ResNet and GoogleNet models while [123], [124] focus the attention on private cloud environments and detection for non-domain experts.

A *Hybrid-based* approach method is proposed in [125] where the framework use more than one complementary filter and a wrapper feature selection approach to identify the most significant runtime behavioural characteristics of malware.

An approach where AI is proliferating is the detection by image visualization. For example, Baptista et al. [126] designed an image-based malware detection tool based on unsupervised learning testing to determine if malicious files could be differentiated from benign ones by focusing on features extracted from their visual representation. In [127], the defined architecture consists of three main components: image generation from malware samples, image augmentation, and classification in a malware family using CNN models. Other similar works are [128], [129], [130], [131]. In the Android world it is worth considering DL-DROID, an automated dynamic analysis framework for Android malware detection. In [132] and [133] satisfying results are obtained using ML and DL techniques. However, the main problem remains the non-Explainability and the subsequent lack of trust in model outcomes, so the next section will explore works that somehow attempt to solve this problem.

3) EXPLAINABLE ARTIFICIAL INTELLIGENCE IN MALWARE DETECTION

One of the main works in this area is Drebin [134]; however, for consistency, it will not be analyzed in-depth as it is a pre-2018 work. Drebin explains his decisions by reporting, for each application, the most influential features, i.e., those present in the application and to which the classifier assigns the highest absolute weights. Melis et al. [135] provide an approach for the Explainability of malware detection in Android systems with an extension of the conceptual approach provided by Drebin on non linear models. Staying focused on Mobile, the authors of [136] use LIME in a method to identify locations deemed important by CNN in the opcode sequence of an Android application to help detect malware, while Kumar et al. [137] propose a static methodology for malware detection in Android where Feature Extraction provides transparency.

XMal [138] is an MLP-based approach with an attention mechanism to detect when an Android App is malware. The interpretation phase aims to automatically produce neural language descriptions to interpret key malicious behaviours within apps. Although the method is not so clear, the authors say they achieve better performance in interpretation than LIME and DREBIN.

The authors in [139] propose a backtracking method to provide a high-fidelity explanation of the DL detection method. The backtracking method selects the most important features

contributing to the classification decision, thus resulting in a transparent and multimodal framework.

Feichtner *et al.* [140] designed a Convolutional Neural Network (CNN) to identify sample-based correlations between parts of the description text and the permission groups an app requests. They employ LIME to calculate a score for each word that shows the output's significance and visualize it as a heatmap.

As analyzed in the previous section, several methods focus on malware detection as an image; in [141], the authors propose a method relying on application representation in terms of images used to input an Explainable Deep Learning model. They represent a mobile application in terms of image and localize the salient parts useful to the model to output a certain precision by exploiting the Grad-CAM algorithm. In this way, the analyst can acquire knowledge about the areas of the image symptomatic of a specific prediction.

Shifting the focus from mobile applications to more general ones, LEMNA [101] is one of the main methods in the landscape of Explainability techniques. It was developed specifically for DL-Based Security Applications and is, therefore, one of the references in the general field of CyberSecurity. It was included in this section because the authors' primary experimentation is conducted on a Malware Detection Dataset. Given a sample of input data, LEMNA generates a small set of interpretable features to explain how the input sample is classified. The central idea is to approximate a local area of the complex DL decision boundary using a simple interpretable model. LEMNA uses a fused lasso-enhanced mixed regression model to generate high-fidelity explanation results for a range of DL models, including RNN.

DENAS [142] is a rule generation approach that extracts knowledge from software-based DNNs. It approximates the nonlinear decision boundary of DNNs, iteratively superimposing a linearized optimization function.

CADE [143] is designed to detect drifting samples that deviate from the original training distribution and provide the corresponding explanations to reason the meaning of the drift. The authors derive explanations based on distance changes, i.e., features that cause the most significant changes to the distance between the drifting sample and its nearest class. It was included in this paragraph because it is tested on a Malware detection dataset.

Pan *et al.* [144], [145] in two related works propose a hardware-assisted malware detection framework developing a regression-based Explainable Machine Learning algorithm. They apply a Decision Tree or Linear Regression to interpret the final result.

In order to understand how a Deep Network architecture generalizes to samples that are not in the training set and explains the outcomes of deep networks in real-world testing, the authors of [146] propose a framework that interpolates between samples of different classes at different layers. By examining the weights and gradients of various levels in the MalConv architecture [147] and figuring out what the architecture discovers by examining raw bytes from the

binary, they try to use this framework to demystify the workings of the MalConv architecture. As a result, they can better explain the workings of ML algorithms and the decisions they make using the proposed framework. Additionally, the analysis will enable network inspection without starting from scratch.

Hsupeng *et al.* [148] introduce an explainable flow-data classification model for hacker attacks and malware detection. The flow data used for training the model is converted from packets by CICFlowMeter. This process significantly shrank the data size, reducing the requirement for data storage. For Explainability, they utilize SHAP further to investigate the relation between cyberattacks and network flow features.

MalDAE [149] is a framework that explores the difference and relation between the dynamic and static API call sequences, which are correlated and fused by semantics mapping. MalDAE provides a practical and explainable framework for detecting and understanding malware based on correlation and fusion of the static and dynamic characteristics. The explainable theoretical framework divides all API calls into several types of malicious behaviours according to their impact on security and builds a hierarchical malware explanation architecture.

Several works in the literature attempt to interpret malware detection by generating *Adversarial* attacks. The authors in [150] discovered that MalConv neural network does not learn any useful characteristics for malware detection from the data and text sections of executable files but instead has a tendency to learn to distinguish between benign and malicious samples based on the characteristics found in the file header. Based on this discovery, they devised a novel attack method that creates adversarial malware binaries by altering a small number of file header bytes. For the explanation, they use Feature Attribution to identify the most influential input features contributing to each decision and adapt it to provide meaningful explanations for classifying malware binaries. Other such works are [151], [152] employing SHAP and [153] proposing a new explanation algorithm to identify the root cause of evasive samples. It identifies the minimum number of features that must be modified to alter the decision of a malware detector, using Action Sequence Minimizer and Feature Interpreter.

To conclude the section, it is necessary to analyze the work of Fan *et al.* [154]. They designed principled guidelines to assess the quality of five explanation approaches by designing three critical quantitative metrics to measure their *Stability*, *Robustness*, and *Effectiveness*. The five explanation approaches are SHAP, LIME, Anchors, LEMNA and LORE. Based on the generated explanation results, they conducted a sanity check of such explanation approaches in terms of the three metrics mentioned. Based on their analysis, the ranking of the five explaining approaches in terms of the *Stability* metric is $LIME \geq SHAP > Anchors > LORE > LEMNA$. The ranking of the five explaining approaches in the *Robustness* metric is $LIME > SHAP > Anchors > LORE > LEMNA$.

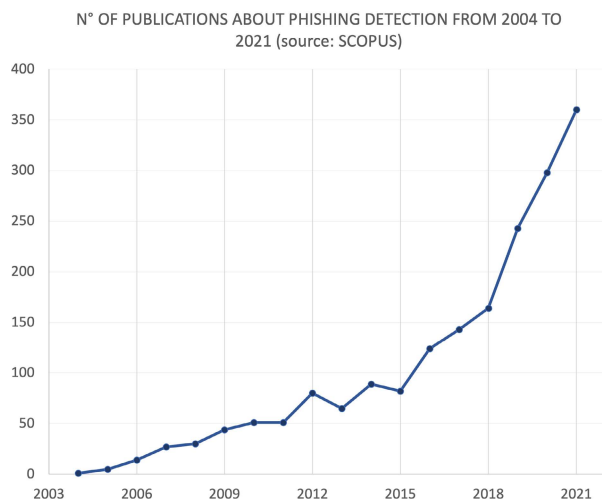


FIGURE 7. Phishing attacks grouped per Quarter⁹.

In the *Effectiveness* metric is $LIME > LORE > Anchors \geq SHAP > LEMNA$.

4) CONSIDERATIONS ABOUT MALWARE DETECTION AND XAI

Several recent publications attempting to explain the results of a malware detector have been reviewed. The significantly smaller number of algorithms that perform detection using images stands out compared to DL, and black-box ML approaches. Another factor to note is the significant effort put into developing Explainable methods in Mobile environments, particularly on Android platforms. Comparing the Black-box and Explainable methods, it is surprising how fewer graph-based methods are used in the latter than in the former; using these for greater transparency might be a good starting point. Several articles use established techniques with *Post-hoc Explainability* that can help the analyst understand the basis on which the model is categorized, particularly SHAP and LIME. Another widely used technique is Feature Attribution, which works similarly to the above approaches. What appears to be obvious is the necessity for applications created with *Intrinsic Explanation* rather than *Post-hoc*, as is usually the case. The Explanation in these cases is built during data training. The model should be a *Hybrid of Signature- and Anomaly-based* methodologies that, when applied together, can give significant benefits. However, it should be recognized that significant progress is being made in this area.

C. PHISHING AND SPAM DETECTION

Phishing refers to a particular type of Internet fraud; the purpose of the malicious attackers, in this circumstance, is to get hold of users' personal and confidential data. More specifically, phishers practice the theft of logins and passwords, credit card and bank account numbers, and additional confidential data.

Spam is also called junk mail. It has existed almost as long as the internet as a means of selling products or services to a larger market of buyers than have ever expressed interest in those products or services. After obtaining the email addresses of a considerable number of individuals, spammers bulk send their offers hundreds or thousands at a time. Spam can be very dangerous if it is part of a phishing attempt.

1) WHY XAI IN PHISHING AND SPAM DETECTION ?

According to the IC3 report,⁸ Phishing (including vishing, SMiShing, and pharming) was the most common threat in the United States in 2020, with 241,342 victims. Following that were nonpayment/non-delivery (108,869 victims), extortion (76,741 victims), personal data breach (45,330 victims), and identity theft (43,330 victims). These data show how huge this problem directly affects the population, which, if not well educated, can easily fall into the trap. The Figure 8 proves the dizzying amount of attention that Phishing attack detection is attracting from academics in recent years. Explaining to a user why a particular email is a phishing attempt or why it has been classified as Spam is no slight advantage. XAI in this field is directly connected to the population that could benefit from it to prevent a threat that is now constant.

2) ARTIFICIAL INTELLIGENCE IN PHISHING AND SPAM DETECTION

Phishing. State of the art on the application of AI in Phishing Detection is substantial, so only recent works with the most significant impact in terms of citations have been analyzed.

Hybrid Ensemble Feature Selection (HEFS) is an interesting approach proposed in [155] with a new feature selection framework. In the first phase of HEFS, a novel Cumulative Distribution Function gradient (CDF-g) algorithm is exploited to produce primary feature subsets, which are then fed into a data perturbation ensemble to yield secondary feature subsets. The second phase derives a set of baseline features from the secondary feature subsets using a function perturbation ensemble. The best performance is achieved with Random Forest. The latter is one of the seven implemented and compared models for the real-time detection of phishing web pages by investigating the URL of the web page explored in [156]. In [157], Yerima *et al.* propose an approach based on a Convolutional Neural Network tested on a dataset obtained from 6,157 genuine and 4,898 phishing websites; a small dataset instead is used in [158] where the authors introduce a Deep Belief Network (DBN). Jain *et al.* propose a ML-based novel Anti-Phishing approach that extracts the features from the client-side only. They examined the various attributes of Phishing and legitimate websites in-depth. As a result, they identified nineteen outstanding features to distinguish Phishing websites from legitimate ones. DTOF-ANN (Decision Tree and Optimal Features based Artificial Neural Network) [159] is a Neural-Network

⁸https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf

PHISHING ATTACKS FROM Q1-2016 TO Q4-2021

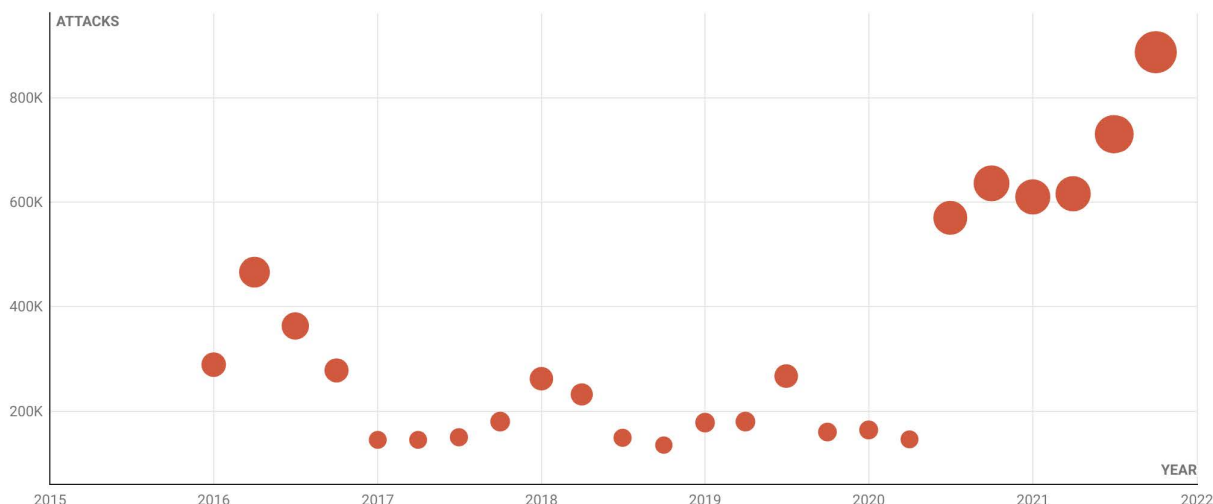


FIGURE 8. Phishing Detection Publications from 2004 to 2021, retrieved from Scopus using as search key [TITLE-ABS-KEY (phishing AND detection)].

Phishing detection model based on a Decision Tree and Optimal feature selection.

The authors of [160] propose Jail-Phish, a *Heuristic* technique which uses Search Engine results and *Similarity-based* features to detect Phishing sites.

The last work to be highlighted for Phishing Detection is PhishBench [161], a benchmarking framework that can help researchers by providing a template to develop new methods and features as well as a platform to compare their proposed techniques with previous works.

Spam. Shifting to Spam Detection, an intelligent system that is based on Genetic Algorithm (GA) and Random Weight Network (RWN) is proposed in [162]. A similar proposal is given by [163] where the authors propose a combination of the Word Embedding technique and Neural Network algorithm.

Barushka *et al.* [164] propose a Spam filter integrating an N-gram tf-idf feature selection, a modified distribution-based balancing algorithm and a regularized Deep multi-layer perceptron NN model with rectified linear units (DBB-RDNN-ReL). In the same wake Douzi *et al.* [165] present a *Hybrid* approach based on the Neural Network model Paragraph Vector-Distributed Memory (PV-DM).

In [166], the authors propose Spam detection in social media with a DL architecture based on Convolutional Neural Network (CNN) and Long Short Term Neural Network (LSTM).

DeepCapture is an image spam email detection tool based on a Convolutional Neural Network (CNN). The key idea is built on a CNN-XGBoost framework consisting of eight layers only with a large number of training samples using data augmentation techniques tailored towards the image Spam

detection task. The evaluation is done on available datasets comprising 6,000 spam and 2,313 non-spam image samples. Other interesting works are [167], [168].

These works are mostly based on Deep Neural Networks in which Interpretability and Explainability of the final detection are challenging, so the next section will analyze the state of the art of explainable models in Phishing and Spam Detection.

3) EXPLAINABLE ARTIFICIAL INTELLIGENCE IN PHISHING AND SPAM DETECTION

The current state of the art for Phishing and Spam detection with explainable methodologies is relatively poor. Therefore, techniques that are not created on-demand for Phishing and Spam Detection but use datasets targeted at these application domains were also considered.

Phishing. Phishpedia [169] is a *Hybrid* DL system that addresses two prominent technical challenges in phishing identification, (i) accurate recognition of identity logos on webpage screenshots and (ii) matching logo variants of the same brand. The authors compare the identity logo and input box providing Explainable annotations on webpage screenshots for the Phishing report.

Two works where the goal is not Phishing detection, but a dataset of this type is used for tests are [170], [171]. The first is based on a Deep embedded Neural Network expert system (DeNNeS) with a rule extraction algorithm for Explainability. The second is based on the Multi-Modal Hierarchical Attention mechanism (MMHAM) that permits the Explainability thanks to the hierarchical system.

Kluge *et al.* [172] propose a framework to convey to the user which words and phrases in an e-mail influenced a Phishing detector's classification of the e-mail as suspicious.

⁹Source: <https://apwg.org/trendsreports/>

They do it by locally perturbing inspiring to Anchors. The last analyzed work is [173], where the authors use LIME and Explainable Boosting Machine (EBM) [174].

Spam. The authors of [175] looked into how different ML explanations, ML model's accuracy, and user confidence in the ML model affect user performance in a simulated Spam detection task. According to their findings, a user's confidence level in the model significantly influences the decision process. Users performed better when using an accurate model. Participants were more likely to spot false alarms generated by the more accurate model and more willing to follow through on a model "miss" when an additional model explanation was given.

FreshGraph [176] is a two-step system for recommending new products to target people that is Spam-aware. First, use item-user Meta-Path similarity and then entropy encoding measurements on a heterogeneous information network structure to identify false positives from candidate lists and avoid potential Spam. The suggested approach takes advantage of the semantic data stored within the graph structure, which considers user activity in addition to item content aspects for more precise audience targeting. Graph structure provides Explainability.

Gu et al. [177] examine the use of DL models to predict the effectiveness of outbound telemarketing for insurance policy loans to decrease Spam problems created by phoning non-potential customers. They propose an Explainable multiple-filter Convolutional Neural Network (XmCNN) to reduce overfitting. Explainability is calculated using feature importance by including a CancelOut layer after the input layer.

These two methods avoid getting into spam and are not spam detector methods. However, they still use Explainable methods of AI to avoid spam; that is why they were analyzed in this section.

The following analysis will focus on techniques that were not created to avoid Spam but instead use Spam datasets as testing. GRACE [178] generates contrastive samples that are concise, informative and faithful to the neural network model's specific prediction. SLISEMAP [179] finds local Explanations for all data items and builds a (typically) two-dimensional global visualization of the black box model such that data items with similar Local Explanations are projected nearby. [180], [181] are two works focused on text classification that use Spam datasets.

4) CONSIDERATIONS ABOUT PHISHING AND SPAM DETECTION AND XAI

As anticipated earlier, state of the art of Explainable Artificial Intelligence in Phishing and Spam detection is very meagre. From the analysis, very few methods are built *Ad-hoc* for detecting these two types of Cyber-attacks. Phishing and Spam are the main threats affecting anyone using a technological device, so using AI for prevention and detection is necessary. AI that simultaneously conveys assurance about the decision made and provides awareness is required to prevent

the decision-making process from becoming less effective for the business and the individual user. As seen in the analysis conducted in [175], the user accepts AI makes mistakes, as long as it is explained how and why so that it can improve in the case of a false negative above all. A consideration beyond XAI in CyberSecurity is the education that must be provided to everyone with a technological device which happens to be surfing the internet where Phishing and Spam are continually around the corner. Similar to how one trains models, one might devise strategies to teach individuals to avoid falling victim to these scams. These strategies need to be Explainable so that anyone can comprehend why certain decisions are taken.

D. BOT (Net) DETECTION

A "Bot" or Robot, is a software program that performs automatic, repetitive, preset operations. Bots often mimic or replace the behaviour of human users. Since they are automated, they work considerably more quickly than actual individuals [182].

Malware and Internet bots can be programmed/hacked to access users' accounts, search the Internet for contact information, transmit Spam, and execute other dangerous operations. Attackers may use malicious Bots in a Botnet, or network of Bots, to launch these attacks and conceal their source. A Botnet is a collection of online-connected devices running one or more Bots, frequently without the owners' knowledge. Since each device has a unique IP address, Botnet activity comprises many IP addresses, making it more challenging to locate and stop the source of malicious Bot traffic. When used to infect additional computers, Spam e-mail recipients' devices can help Botnets grow larger. They are commanded by hackers known as Botmasters or Bot herders.

Botnets are hard to spot since they consume very few computer resources. This keeps them from interfering with applications' regular operation and does not make the user suspicious. However, the most sophisticated Botnets can also alter their behaviour by the CyberSecurity systems of the PCs to evade detection. Most of the time, users are unaware that their devices are part of a Botnet and are under the control of online criminals [183].

1) WHY XAI IN BOT (Net) DETECTION?

Spamhaus monitors both IP addresses and domain names used by threat actors to run botnet Command & Control (C&C) servers. As a result, Spamhaus Malware Labs researchers found and blacklisted 17, 602 botnet C&C servers hosted on 1, 210 distinct networks.¹⁰ This represents a massive 71.5% increase over the number of botnet C&Cs witnessed in 2018. Since 2017, the number of newly discovered botnet C&Cs has nearly doubled, rising from 9, 500 to 17, 602 in 2019. The figure 9 shows the increasing attention of researchers in this area.

¹⁰<https://www.spamhaus.org/news/article/793/spamhaus-botnet-threat-report-2019>

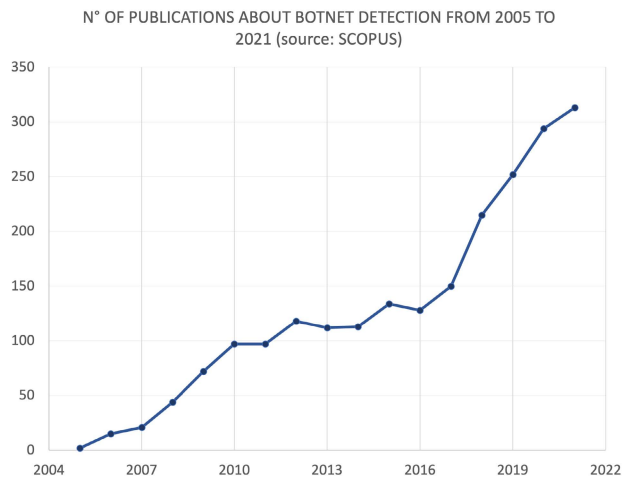


FIGURE 9. BotNet Detection Publications from 2005 to 2021, retrieved from Scopus using as search key [TITLE-ABS-KEY (botnet AND detection)].

AI, applied with Explainable methods, is certainly among the best methods to counter this phenomenon in which a huge number of resources have to be vanquished.

2) ARTIFICIAL INTELLIGENCE IN BOT (Net) DETECTION

This section quickly reviews the newest and most cited methods in BotNet Detection. For Bot Detection, refer to the comprehensive survey by Cresci *et al.* [182].

Fast-flux hunter (FFH) [184] is a framework that can improve the performance level in detecting and predicting unknown and Zero-day fast-flux Botnets. FFH distinguishes the fast-flux Botnets domain from legitimate domains in an online mode based on new rules, features, or classes to enhance learning using the EFuNN algorithm.

TS-ASRCAPS [185] is a framework based on double-stream networks, which uses multimodal information to reflect the characteristics of Domain Generation Algorithms, and an attention-sliced recurrent neural network to automatically mine the underlying semantics.

The authors of [186] propose a memory-efficient DL method, named LS-DRNN, for Botnets attack detection in IoT networks. S-DRNN method employs SMOTE and DRNN algorithms only. However, LS-DRNN combines Long Short-Term Memory Autoencoder (LAE), SMOTE, and DRNN algorithms.

The framework proposed in [187] uses ML combined with a honeynet-based detection method for predicting if an IoT device can be a part of a Botnet.

In [188], the authors use a CNN to perceive subtle differences in power consumption and detect Anomalies.

In [189], the authors point out one of their proposal's main cons, the framework's non-Explainability. They emphasize that this is a problem with DL models and that this implies a lack of confidence. The following section will analyze frameworks that try to explain why a particular classification is made. Other interesting works are [190], [191].

3) EXPLAINABLE ARTIFICIAL INTELLIGENCE IN BOT (Net) DETECTION

BotStop [192] is a *Packet-based* Botnet detection system that examines incoming and outgoing network traffic in an IoT device to prevent infections from Botnets. The proposed system is founded on Explainable ML algorithms thanks to SHAP use with features extracted from network packets. Once an attack is detected, the source is blocked. Always SHAP is used in [193] to determine the relevant traffic features in a framework to detect traffic generated by a Bot and then determine the type of Bots using a Convolutional Neural Network.

Suryotrisongko *et al.* [194] propose the XAI and OSINT combination for Cyber Threat Intelligence Sharing in preventing Botnet DGA. This research applied four existing XAI techniques: Anchors, SHAP, Counterfactual Explanation and LIME. This latter is also used in [195] and [196] where the final goal is the detection in IoT Networks.

BD-GNNExplainer [197] is a Botnet Detection Model based on Graph Neural Network. The explanation is attributable to subgraph decomposition theory [198], where it is feasible to determine whether the learned model is interpretable by identifying the subgraph with the most significant influence on prediction and judging whether the subgraph is faithful to general knowledge.

Reference [199], [200], [201], three explainable studies focused on DGA-based botnet detection, are also worth mentioning, as is [202], in which the authors created a Gradient-based Explainable Variational Autoencoder for *Network Anomaly Detection* utilizing a BotNet dataset as a test.

Bot-Detective [203] is an explainable Twitter bot detection service with crowdsourcing functionalities that uses LIME. LIME is also used in JITBot [204], An Explainable Just-In-Time Defect Prediction Bot, and in [205], a bot-type classification schema.

SHAP and LIME are used in [206] for game BOT detection, while in [207], the authors used a Decision Tree model, Explainable by definition, for automatic detection on Twitter with a particular case study on posts about COVID-19.

4) CONSIDERATIONS ABOUT BOT (Net) DETECTION AND XAI

As noted in the previous sections, almost all of the frameworks declared Explainable use existing methods for *Post-hoc* Explanation, SHAP and LIME above all. In BotNet Detection, the almost total focus on IoT networks and devices should be especially noted, demonstrating that these occupy a very important slice of the Net. As in the case of Spam and Phishing, it is critical to alert if you have entered a BotNet and are feeding it unknowingly, and even more important to Explain what you have inferred and how you got into it, so that you can avoid falling into it again in the future. It is moving in this direction, as evidenced by the increasing number of publications on the subject, however, one must consider that also improving is the malicious part of the

fight. That is why it is increasingly important that supporting human decisions is AI, which can counter a considerable part of these attacks in an automated way. For there to be the right cooperation between human and AI, Explainability of the latter is necessary to build trust in the former.

E. OTHER CYBERSECURITY TREATS

The Macro Categories considered up to this point are those in which the greatest effort has been spent with the purpose of applying Explaining Artificial Intelligence in CyberSecurity.

Fraud Detection. The financial sector is one of the ones most frequently targeted by cyberattacks. Frauds are frequent Cyber-attacks linked to money and reputation issues in this field. Data leaks and illegal credit losses may be the root of such attacks.

xFraud, an Explainable fraud transaction detection framework based on Graph Neural Networks (GNN), is presented in [208]. The authors designed a *Learnable Hybrid Explainer* that leverages GNNExplainer and centrality measures to learn node- and edge-level Explanations simultaneously.

Srinath et al. [209] present an Explainable Machine Learning framework for identifying credit card defaulters using DALEX [210].

Zero-Day Vulnerabilities. The term “Zero-day” refers to recently identified security flaws that hackers utilize to attack systems. The expression “Zero-day” alludes to the notion that the vendor or developer has “Zero days” to repair the defect because they have just become aware of it. When hackers use a vulnerability before developers have a chance to fix it, a Zero-day assault is launched.

The authors of [211] propose a new visualization technique using similarity matrices of features depicting behaviour patterns of malware and displaying them in image form for faster analysis for detection of Zero-day malware. Kumar et al. [212] use Shapley Ensemble Boosting and Bagging Approach instead for the same goal.

The authors in [213] propose a method for Zero-Day Web Attacks delivering outlier explanations. The method shows that Explanations can be backwards transformed through n-gram encoding and dimensionality reduction.

In [214], Zhou et al. define a Zero-day artificial immune system driven by XAI for intrusion detection in telecommunications. The central part of the artificial immune system is extracting strict rules for benign traffic. It uses a Decision Tree that is, by definition, a white-box model.

Digital Forensics. Digital Forensics or Computer Forensics finds its place in Forensic Science or Criminalistics. It is, therefore, that branch of Forensic science that deals with investigating the contents of digital devices, during investigation and trial, for evidentiary purposes. The collected data are identified, acquired, analyzed, and a technical report is written.

Hall et al. [215] assert that the application of AI in digital/network forensics is still a “Black box” at this time, requiring verification by digital/network Forensic investigators, and is therefore unlikely to be justified in court.

Furthermore, the admissibility of digital/network analysis performed by XAI in court is still debatable as it would necessitate a review of applicable laws (e.g., evidence law). However, XAI can be used efficiently and legally in the future to support the digital/network forensic profession if it is not viewed as a replacement for a digital/network forensic examiner but rather as a reliable tool to aid in investigations.

ATLE2FC [216] is a model for IoT Forensics using Ensemble Classification with an Explainable layer consisting of FPGrowth with GRU-based RNN classifier for rule estimation and severity classification.

For media forensic investigations focusing on media forensic object modification detection, such as DeepFake detection, a domain-adapted forensic data model is introduced in [217] and [218].

Cyber Physical Systems. When an adversary gains access to a computer system that controls equipment in a manufacturing facility, oil pipeline, refinery, electric generating plant, or other similar infrastructure, they can control the operations of that equipment to harm those assets or other property. This is known as a Cyber-Physical attack on critical infrastructure. Cyber-Physical attacks pose a risk not only to the owners and operators of those assets but also to their suppliers, clients, enterprises, and people nearby the targeted asset, as well as to any individual or entity they could negatively impact. For example, a Cyber-Physical attacker may take down cameras, switch off the lights in a building, cause a car to wander off the road, or make a drone land in the hands of adversaries.

Wickramasinghe et al. [219] propose a Desiderata on Explainability of unsupervised approaches in Cyber-Physical Systems since they generate a large amount of unlabeled data. These are potential solutions for meaningfully mining these data, maintaining and improving desired functions, and improving the safety of these systems.

An Explainable Cyber-Physical Systems based on Knowledge Graph is proposed in [220] for Energy Systems while in [221] the authors propose a framework to build Self-Explainable Cyber-Physical System.

Crypto-Jacking. Crypto-jacking, a new Malware that resides on a computer or mobile device and uses its resources to “mine” Cryptocurrencies, is a severe online threat. In addition to compromising various devices, including PCs, laptops, cellphones, and even network servers, Crypto-Jacking can take control of web browsers. Using Crypto-Jacking, criminals compete with sophisticated Crypto mining operations without the high overhead costs by stealing computational power from victims’ devices.

It is a threat comparable to BotNets, where unknowingly the user feeds activities with malicious purposes through their device.

There are no works that make Explainable Artificial Intelligence methods in the detection of Cryptojacking, one that goes in this direction in the detection of Cryptomining is that of Karn et al. [222]. They designed and implemented an automated cryptomining pod (management of applications inside containers) detection in a

TABLE 2. Summary of methods.

APPLICATION	REF.	YEAR	AI MODEL	EXPLANATION METHOD	SUMMARY
INTRUSION DETECTION SYSTEMS	[90]	2018	LINEAR CLASSIFIER/MLP	ADVERSARIAL	H-L-AG
	[102]	2018	MLP	KEY FEATURE EXTRACTION	H-L-AG
	[100]	2019	RNN	REGRESSION + FUSED LASSO	I-L-SP
	[104]	2019	SVM	LIME	H-L-AG
	[89]	2020	ANN/DT	CLUSTERING	H-L-AG
	[106]	2020	RF/NB/XGB	SHAP	H-L (G) -AG
	[91], [92]	2020, 2021	MULTICLASS/BINARY CLASSIFIERS	SHAP	H-L (G) -AG
	[94]	2021	RF	SHAP	H-L (G) -AG
	[95]	2021	SVM	LIME	H-L-AG
	[96]	2021	BRCG + LOGRR + PROTODASH+ CEM	BRCG + LOGRR + PROTODASH+ CEM	H-L (G) -AG
	[107]	2021	DNN	SHAP/BRCG/LIME/CEM/PROTODASH	H-L (G) -AG
	[108]	2021	DFE/RF	SHAP	H-L-AG
	[88]	2022	DT	WHITE-BOX MODEL	I-G-SP
	[93]	2022	RF	SHAP	H-L (G) -AG
[103]	2022	RF	SHAP	H-L (G) -AG	
[105]	2022	CNN	GRAD-CAM	H-G-SP	
MALWARE DETECTION	[135]	2018	SVM-RBF/RF	GRADIENT-BASED	H-L (G) -AG
	[137]	2018	SVM	KEY FEATURE EXTRACTION	H-L-AG
	[101]	2018	RNN/MLP	REGRESSION + FUSED LASSO	I-L-SP
	[139]	2019	CNN	BACKTRACKING	H-L-SP
	[150]	2019	MALCONV	FEATURE ATTRIBUTION	H-L-AG
	[140]	2020	CNN	LIME	H-L-AG
	[142]	2020	DNN	RULE-BASED	H-G-AG
	[144], [145]	2020, 2022	RNN/DT	LINEAR REGRESSION/WHITE-BOX MODEL	H-L-SP/I-G-SP
	[136]	2021	CNN	LIME	H-L-SP
	[138]	2021	MLP	KEY FEATURE EXTRACTION	H-L-AG
	[141]	2021	CNN	GRAD-CAM	H-L-AG
	[143]	2021	MLP	CLUSTERING	I-L-AG
	[148]	2022	XGB	SHAP	H-L-AG
	PHISHING DETECTION	[169]	2021	FASTER-RCNN	COSINE SIMILARITY
[172]		2021	RNN	Similar to ANCHORS	H-L-AG
[173]		2021	RF/SVM/EBM	LIME/EBM	H-L (G) -AG
SPAM DETECTION	[177]	2021	CNN	XMCNN filter	H-L-SP
BOTNET DETECTION	[196]	2019	RF/KNN/DECISION TREE	LIME	H-L-AG
	[195]	2021	RF/EXTRA TREES	LIME	H-L (G) -AG
	[192]	2022	XGB	SHAP	H-L (G) -AG
	[193]	2022	CNN	SHAP	H-G-AG
	[194]	2022	NB/LR/RF/EXTRA TREES	SHAP/LIME/ANCHORS/COUNTERFACTUAL EXPLANATION	H-L (G) -AG
[197]	2022	GNN	SUBGRAPH DECOMPOSITION	I-L-SP	
FRAUD DETECTION	[208]	2021	GNN	CENTRALITY MEASURES	I-L-SP
	[209]	2022	XGB	DALEX	H-L-AG
ZERO-DAY VULNERABILITIES	[211]	2018	CNN	SIMILARITY MATRICES	H-G-AH
	[212]	2022	RF/XGB/EXTRA TREES	SHAPLEY VALUES	H-L-AG
	[214]	2022	DT	WHITE-BOX MODEL	I-G-SP
DIGITAL FORENSICS	[216]	2021	RNN	FPGROWHT	I-L-SP
CYBER PHYSICAL SYSTEMS	[220]	2021	KNOWLEDGE GRAPH	EXPLANATION GENERATION ALGORITHM	H-L-SP
CRYPTO MINING	[222]	2020	RNN	SHAP/LIME	H-L-AG

Legend Summary: I: Intrinsic, H: Post-hoc, G: Global, L: Local, SP: Model-specific, AG: Model-agnostic

Kubernetes cluster. Explainability is provided using SHAP, LIME, and a novel auto-encoding-based scheme for LSTM models.

VI. DISCUSSION AND CHALLENGES

Due to the broad spectrum of XAI approaches, analyzing the different surveys involving these works were preferred to

better orient the reader. It is also unthinkable to include all studied papers; hence only a selection of works was discussed in this survey for synthesis and relevancy considerations, prioritizing all work that proposed XAI methods with application in CyberSecurity.

Table 2 summarizes the principal works of XAI for each CyberSecurity application analyzed with a focus on the

TABLE 3. Summary of most used cyber dataset in main cyber application fields.

APPLICATION	NAME	YEAR	REF.	URL
INTRUSION DETECTION SYSTEMS	KDDCUP 99	1999	[223]	http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
	NSL-KDD	1999	[224]	https://www.unb.ca/cic/datasets/nsl.html
	ADFA-WD	2013	[225]	https://research.unsw.edu.au/projects/adfa-ids-datasets
	ADFA-LD	2014	[226]	https://research.unsw.edu.au/projects/adfa-ids-datasets
	UNSW-NB15	2015	[227]	https://research.unsw.edu.au/projects/unsw-nb15-dataset
	AWID2	2016	[228]	https://icsdweb.aegean.gr/awid/
	CICIDS2017	2017	[229]	https://www.unb.ca/cic/datasets/ids-2017.html
	CSE-CIC-IDS2018	2018	[230]	https://www.unb.ca/cic/datasets/ids-2018.html
MALWARE DETECTION	BIG 2015	2015	[231]	ronen2018microsoft
	EMBER	2017	[232]	https://github.com/elastic/ember
	MALREC	2018	[233]	https://giantpanda.gtisc.gatech.edu/malrec/dataset/
	MICROSOFT MALWARE PREDICTION	2018	[234]	https://github.com/greulist137/Microsoft-Malware-Prediction
	CIC-INVESANDMAL2019	2019	[235]	https://www.unb.ca/cic/datasets/invesandmal2019.html
SPAM/PHISHING DETECTION	LING SPAM	2000	[236]	https://www.kaggle.com/datasets/mandygu/lingspam-dataset
	ENRON	2004	[237]	https://www.cs.cmu.edu/~enron/
	SPAM ASSASSIN	2004	[238]	https://www.kaggle.com/datasets/beatoa/spamassassin-public-corpus
BOTNET DETECTION	ISOT	2013	[239]	https://www.unb.ca/cic/datasets/botnet.html
	UMUDGA	2020	[240]	https://data.mendeley.com/datasets/y8ph45msv8/1

ML/DL model, the type of explanation and a summary concerning the taxonomy presented in section II-A.

Table 3 presents the main datasets for each application field encountered during the survey, highlighting the use of aged datasets. Methods and datasets are ordered by year for each application field.

The selection criteria were based mainly on a backward and forward snowballing strategy that consists of using the reference list of the selected papers and the citations to these papers to identify additional papers [241]. The proposed review was founded on a solid foundation that included the most critical areas of XAI and CyberSecurity subjects. Because of the investigated domains' importance and rapid growth, it has been determined that non-traditional sources are also necessary to analyze since they are essential and impactful in the field. In the following the main challenges emerged after the review conducted.

More formalism is needed. XAI is a multidimensional target that a single theoretical approach cannot achieve. However, the synergistic employment of techniques from diverse study horizons must be done in a well-integrated manner. In other words, for the area to advance, it needs to be supported by a separate research community, which, at this point of development, should primarily focus on increased formalism. The reference is mainly to works that apply Explainable Artificial Intelligence methods in CyberSecurity without specifying in what and how, at what level, with output reported to whom (whether users, analysts or developers) and especially with what techniques. In the same field of application (e.g., Malware Detection), it would be good to unify the work in terms of Explainability so that those in charge of analyzing and preventing cyber-attacks can have a unified and more understandable view.

Human in the loop. It is not enough to explain the model; the user must comprehend it. Furthermore, even with an appropriate explanation, establishing such an understanding may necessitate supplementary responses to queries that

users are likely to ask. Thus, explainability can only occur through human-machine interaction. In [242], the authors present an example and approach for creating a concept for an XAI-driven junior cyber analyst based on understanding the information needs of both humans and AI components in terms of the work context and workflow. This method may be required to design future systems that people can use, particularly for critical systems where human stakeholders cannot interact with black-box outputs from intelligent agents, as is the case in many CyberSecurity applications. Therefore, the idea and proposal are to think about and build frameworks that have human-machine interaction at their core for CyberSecurity applications, which is vital in many cases. The only way to get there is to build models understandable to humans.

How to achieve Explainability. In the current state of the art, as shown in the Table 2, the proposed methods use post-hoc explanation in most cases. Developing models that provide an intrinsic explanation is a priority; an explanation method developed ad-hoc for that particular type of application is necessary for a field such as CyberSecurity, where one risks providing an assist to the attacker. Moreover, the problem may be precisely in terms of explanation, and the risk is to provide an untruthful output. As pointed out several times in [101], LIME, one of the most widely used methods, assumes that the decision boundary is locally linear. However, when the local decision boundary is non-linear, as it is in the majority of complex networks, those explanation approaches cause significant inaccuracies. In some cases, the linear portion is severely constrained to a relatively tiny region. The artificial data points beyond the linear zone are easily struck by standard sampling methods, making it hard for a linear model to estimate the decision boundary near x . The challenge then is not easy, the inverse correlation between model opacity and performance is well known, but an effort is needed to develop increasingly high-performing but transparent models.

Adversarial Attacks. An in-depth investigation of how pattern explanations can provide new attack surfaces for the underlying systems is needed. A motivated attacker can use the information offered by the explanations to perform membership inference and pattern mining attacks, damaging overall system privacy. Regular adversarial attacks are predicated on the assumption that an adversary may inject a perturbation into an input sample that is undetectable to humans, and, as a result, the ground-truth class of the perturbed input does not change. The second issue is that a ML model's projected class changes. Attackers have developed several techniques to exploit weaknesses in XAI-enabled CyberSecurity frameworks. Adversary attacks circumvent authentication systems, such as the XAI-enabled facial authentication system, while poisoning attacks were used to alter or damage training data [243]. To combat these attacks, a solution could be to analyze "Desiderata for adversarial attacks in different scenarios involving explainable ML models" presented in [244].

VII. CONCLUSION

XAI is a framework to help understand and interpret the predictions of AI algorithms. CyberSecurity is an area where AI can analyze datasets and track a wide range of security threats and malicious behaviors. The only way to address the many CyberSecurity challenges, with an increasing number of attacks, is through the integration of human and AI. This paper reviews work proposed in the past five years that seeks to bridge human and machine through explainability. After a careful analysis of the two ecosystems, XAI and CyberSecurity, an analysis was conducted of the areas of CyberSecurity most affected by the use of AI. What distinguishes this work is the exploration of how each method provides explainability for different application areas, highlighting the lack of formalism and the need to move toward a standard. The final analysis explored the most relevant problems and open challenges. Considerable effort is needed to ensure that ad hoc frameworks and models are built for safety and not the application of general models for post-hoc explanation.

REFERENCES

- [1] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nature Mach. Intell.*, vol. 1, no. 12, pp. 557–560, Dec. 2019.
- [2] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019.
- [3] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, "Four principles of explainable artificial intelligence," NIST Interagency, Gaithersburg, MD, USA, Internal Rep. NISTIR-8312, Aug. 2020, doi: [10.6028/NIST.IR.8312](https://doi.org/10.6028/NIST.IR.8312).
- [4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
- [5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
- [6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Apr. 2018, pp. 1–9.
- [7] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," 2018, *arXiv:1805.10820*.
- [8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [9] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.
- [10] S. Morgan. (2020). Special report: Cyberwarfare in the C-suite, online. Cybercrime Magazine. [Online]. Available: <https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/>
- [11] *Enisa Threat Landscape 2020—List of Top 15 Threats*, ENISA, Athens, Greece, 2020.
- [12] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.
- [13] A. Rawal, J. McCoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives," *IEEE Trans. Artif. Intell.*, no. 4, Aug. 2021, doi: [10.1109/TAI.2021.3133846](https://doi.org/10.1109/TAI.2021.3133846).
- [14] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020.
- [15] A. Kale, T. Nguyen, F. C. Harris, Jr., C. Li, J. Zhang, and X. Ma, "Provenance documentation to enable explainable and trustworthy AI: A literature review," *Data Intell.*, pp. 1–41, Feb. 2022, doi: [10.1162/dint_a_00119](https://doi.org/10.1162/dint_a_00119).
- [16] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- [17] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–18.
- [18] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, 2018.
- [19] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8827–8836.
- [20] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 11, no. 5, p. e1424, 2021.
- [21] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
- [22] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.
- [23] G. Riccardo, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.
- [24] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Appl. Sci.*, vol. 12, no. 3, p. 1353, Jan. 2022.
- [25] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Oct. 2021.
- [26] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in AI," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 279–288.
- [27] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2018.
- [28] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.

- [29] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 44, pp. 22071–22080, 2019.
- [30] O. Loyola-Gonzalez, "Black-box vs. White-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.
- [31] G. Vilone and L. Longo, "Explainable artificial intelligence: A systematic review," 2020, *arXiv:2006.00093*.
- [32] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 615–661, Aug. 2021.
- [33] D. V. Carvalho, M. E. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.
- [34] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sasing, and K. Baum, "What do we want from explainable artificial intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artif. Intell.*, vol. 296, Jul. 2021, Art. no. 103473.
- [35] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 648–657.
- [36] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable artificial intelligence for tabular data: A survey," *IEEE Access*, vol. 9, pp. 135392–135422, 2021.
- [37] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.
- [38] M. Z. Siddiqui, S. Yadav, and M. S. Husain, "Application of artificial intelligence in fighting against cyber crimes: A review," *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 2, pp. 118–122, 2018.
- [39] Z. I. Khisamova, I. R. Begishev, and E. L. Sidorenko, "Artificial intelligence and problems of ensuring cyber security," *Int. J. Cyber Criminol.*, vol. 13, no. 2, pp. 564–577, 2019.
- [40] I. A. Mohammed, "Artificial intelligence for cybersecurity: A systematic mapping of literature," *Artif. Intell.*, vol. 7, no. 9, pp. 1–5, 2020.
- [41] H. Suryotrisongko and Y. Musashi, "Review of cybersecurity research topics, taxonomy and challenges: Interdisciplinary perspective," in *Proc. IEEE 12th Conf. Service-Oriented Comput. Appl. (SOCA)*, Nov. 2019, pp. 162–167.
- [42] J.-H. Li, "Cyber security meets artificial intelligence: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 12, pp. 1462–1474, Dec. 2018.
- [43] T. C. Truong, Q. B. Diep, and I. Zelinka, "Artificial intelligence in the cyber domain: Offense and defense," *Symmetry*, vol. 12, no. 3, p. 410, Mar. 2020.
- [44] C. V. Dalave and T. Dalave, "A review on artificial intelligence in cyber security," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, 2022, pp. 304–309.
- [45] M. Akhtar and T. Feng, "An overview of the applications of artificial intelligence in cybersecurity," *EAI Endorsed Trans. Creative Technol.*, vol. 8, no. 29, Dec. 2021, Art. no. 172218.
- [46] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222310–222354, 2020.
- [47] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "AI-driven cybersecurity: An overview, security intelligence modeling and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 3, pp. 1–18, May 2021.
- [48] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: An overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, pp. 1–29, Dec. 2020.
- [49] C. Gupta, I. Johri, K. Srinivasan, Y.-C. Hu, S. M. Qaisar, and K.-Y. Huang, "A systematic review on machine learning and deep learning models for electronic information security in mobile networks," *Sensors*, vol. 22, no. 5, p. 2017, Mar. 2022.
- [50] A. F. Jahwar and S. Y. Ameen, "A review on cybersecurity based on machine learning and deep learning algorithms," *J. Soft Comput. Data Mining*, vol. 2, no. 2, pp. 14–25, Oct. 2021.
- [51] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.
- [52] K. Shaukat, S. Luo, V. Varadharajan, I. Hameed, S. Chen, D. Liu, and J. Li, "Performance comparison and current challenges of using machine learning techniques in cybersecurity," *Energies*, vol. 13, no. 10, p. 2509, May 2020.
- [53] M. Alabadi and Y. Celik, "Anomaly detection for cyber-security based on convolution neural network : A survey," in *Proc. Int. Congr. Hum.-Comput. Interact., Optim. Robotic Appl. (HORA)*, Jun. 2020, pp. 1–14.
- [54] S. Kim and K.-J. Park, "A survey on machine-learning based security design for cyber-physical systems," *Appl. Sci.*, vol. 11, no. 12, p. 5458, Jun. 2021.
- [55] D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A survey of deep learning methods for cyber security," *Information*, vol. 10, no. 4, p. 122, 2019.
- [56] D. Gumusbas, T. Yldrm, A. Genovese, and F. Scotti, "A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems," *IEEE Syst. J.*, vol. 15, no. 2, pp. 1717–1731, Jun. 2021.
- [57] O. Lifandali and N. Abghour, "Deep learning methods applied to intrusion detection: Survey, taxonomy and challenges," in *Proc. Int. Conf. Decis. Aid Sci. Appl. (DASA)*, Dec. 2021, pp. 1035–1044.
- [58] J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, and Y. Xiang, "Deep learning based attack detection for cyber-physical system cybersecurity: A survey," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 3, pp. 377–391, Mar. 2022.
- [59] M. A. Ferrag, L. Maglaras, S. Moschoviannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *J. Inf. Secur. Appl.*, vol. 50, Feb. 2020, Art. no. 102419.
- [60] S. Gamage and J. Samarabandu, "Deep learning methods in network intrusion detection: A survey and an objective comparison," *J. Netw. Comput. Appl.*, vol. 169, Nov. 2020, Art. no. 102767.
- [61] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 139–154, Jan. 2021.
- [62] T. R. Reshmi, "Information security breaches due to ransomware attacks—A systematic literature review," *Int. J. Inf. Manage. Data Insights*, vol. 1, no. 2, Nov. 2021, Art. no. 100013.
- [63] S. Hariharan, A. Velicheti, A. S. Anagha, C. Thomas, and N. Balakrishnan, "Explainable artificial intelligence in cybersecurity: A brief review," in *Proc. 4th Int. Conf. Secur. Privacy (ISEA-ISAP)*, Oct. 2021, pp. 1–12.
- [64] G. Srivastava, R. H. Jhaveri, S. Bhattacharya, S. Pandya, P. K. R. Maddikunta, G. Yenduri, J. G. Hall, M. Alazab, and T. R. Gadekallu, "XAI for cybersecurity: State of the art, challenges, open issues and future directions," 2022, *arXiv:2206.03585*.
- [65] J. N. Paredes, J. Carlos, L. Teze, G. I. Simari, and M. V. Martinez, "On the importance of domain-specific explanations in AI-based cybersecurity systems (technical report)," 2021, *arXiv:2108.02006*.
- [66] L. Viganò and D. Magazzeni, "Explainable security," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS PW)*, Sep. 2020, pp. 293–300.
- [67] V. Ravi et al., "Deep learning for cyber security applications: A comprehensive survey," *TechRxiv*, 2021, doi: [10.36227/techrxiv.16748161.v1](https://doi.org/10.36227/techrxiv.16748161.v1).
- [68] Y. Liu, C. Tantithamthavorn, L. Li, and Y. Liu, "Deep learning for Android malware defenses: A systematic literature review," 2021, *arXiv:2103.05292*.
- [69] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, p. e4150, Jan. 2021.
- [70] S. K. Biswas, "Intrusion detection using machine learning: A comparison study," *Int. J. pure Appl. Math.*, vol. 118, no. 19, pp. 101–114, 2018.
- [71] A. Chawla, B. Lee, S. Fallon, and P. Jacob, "Host based intrusion detection system with combined CNN/RNN model," in *ECML PKDD 2018 Workshops. ECML PKDD 2018 (Lecture Notes in Computer Science)*, vol. 11329, C. Alzate, A. Monreale, H. Assem, A. Bifet, T. S. Buda, B. Caglayan, B. Drury, E. García-Martín, R. Gavaldà, I. Koprinska, S. Kramer, N. Lavesson, M. Madden, I. Molloy, M.-I. Nicolae, and M. Sinn, Eds. Cham, Switzerland: Springer, 2019, doi: [10.1007/978-3-030-13453-2_12](https://doi.org/10.1007/978-3-030-13453-2_12).
- [72] J. Byrnes, T. Hoang, N. N. Mehta, and Y. Cheng, "A modern implementation of system call sequence based host-based intrusion detection systems," in *Proc. 2nd IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Oct. 2020, pp. 218–225.

- [73] R. Gassais, N. Ezzati-Jivan, J. M. Fernandez, D. Aloise, and M. R. Dagenais, "Multi-level host-based intrusion detection system for Internet of Things," *J. Cloud Comput.*, vol. 9, no. 1, pp. 1–16, Dec. 2020.
- [74] E. Besharati, M. Naderan, and E. Namjoo, "LR-HIDS: Logistic regression host-based intrusion detection system for cloud environments," *J. Ambient Intell. Humanized Comput.*, vol. 10, no. 9, pp. 3669–3692, Sep. 2019.
- [75] M. Liu, Z. Xue, X. He, and J. Chen, "SCADS: A scalable approach using spark in cloud for host-based intrusion detection system with system calls," 2021, *arXiv:2109.11821*.
- [76] D. Park, S. Kim, H. Kwon, D. Shin, and D. Shin, "Host-based intrusion detection model using Siamese network," *IEEE Access*, vol. 9, pp. 76614–76623, 2021.
- [77] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 41–50, Feb. 2018.
- [78] Y. Jia, M. Wang, and Y. Wang, "Network intrusion detection algorithm based on deep neural network," *IET Inf. Secur.*, vol. 13, no. 1, pp. 48–53, Jan. 2019.
- [79] M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, "Deep learning approach combining sparse autoencoder with SVM for network intrusion detection," *IEEE Access*, vol. 6, pp. 52843–52856, 2018.
- [80] M. H. Ali, B. A. D. Al Mohammed, A. Ismail, and M. F. Zolkipli, "A new intrusion detection system based on fast learning network and particle swarm optimization," *IEEE Access*, vol. 6, pp. 20255–20261, 2018.
- [81] Z. Wang, "Deep learning-based intrusion detection with adversaries," *IEEE Access*, vol. 6, pp. 38367–38384, 2018.
- [82] B. Yan and G. Han, "Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system," *IEEE Access*, vol. 6, pp. 41238–41248, 2018.
- [83] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network intrusion detection combined hybrid sampling with deep hierarchical network," *IEEE Access*, vol. 8, pp. 32464–32476, 2022.
- [84] Y. Yu and N. Bian, "An intrusion detection method using few-shot learning," *IEEE Access*, vol. 8, pp. 49730–49740, 2020.
- [85] Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization," *IEEE Access*, vol. 8, pp. 42169–42184, 2020.
- [86] J. Clements, Y. Yang, A. A. Sharma, H. Hu, and Y. Lao, "Rallying adversarial techniques against deep learning for network security," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2021, pp. 01–08.
- [87] G. Andresini, A. Appice, N. Di Mauro, C. Loglisci, and D. Malerba, "Multi-channel deep feature learning for intrusion detection," *IEEE Access*, vol. 8, pp. 53346–53359, 2020.
- [88] T. Dias, N. Oliveira, N. Sousa, I. Praça, and O. Sousa, "A hybrid approach for an interpretable and explainable intrusion detection system," in *Intelligent Systems Design and Applications. ISDA 2021 (Lecture Notes in Networks and Systems)*, vol. 418, A. Abraham, N. Gandhi, T. Hanne, T. P. Hong, T. N. Rios, and W. Ding, Eds. Cham, Switzerland: Springer, 2022, doi: [10.1007/978-3-030-96308-8_96](https://doi.org/10.1007/978-3-030-96308-8_96).
- [89] M. Szczepanski, M. Choras, M. Pawlicki, and R. Kozik, "Achieving explainability of intrusion detection system by hybrid oracle-explainer approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.
- [90] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2018, pp. 3237–3243.
- [91] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.
- [92] Y. Wang, P. Wang, Z. Wang, and M. Cao, "An explainable intrusion detection system," in *Proc. IEEE 23rd Int. Conf. High Perform. Comput. Commun., 7th Int. Conf. Data Sci. Syst., 19th Int. Conf. Smart City, 7th Int. Conf. Dependability Sensor, Cloud Big Data Syst. Appl. (HPCC/DSS/SmartCity/DependSys)*, Dec. 2021, pp. 1657–1662.
- [93] T.-T.-H. Le, H. Kim, H. Kang, and H. Kim, "Classification and explanation for intrusion detection system based on ensemble trees and SHAP method," *Sensors*, vol. 22, no. 3, p. 1154, Feb. 2022.
- [94] S. Wali and I. Khan, "Explainable AI and random forest based reliable intrusion detection system," TechRxiv, 2021, doi: [10.36227/techrxiv.17169080.v1](https://doi.org/10.36227/techrxiv.17169080.v1).
- [95] E. Tcydenova, T. W. Kim, C. Lee, and J. H. Park, "Detection of adversarial attacks in ai-based intrusion detection systems using explainable AI," *Hum.-Centric Comput. Inf. Sci.*, vol. 11, pp. 1–14, Sep. 2021.
- [96] H. Liu, C. Zhong, A. Alnusair, and S. R. Islam, "FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques," *J. Netw. Syst. Manage.*, vol. 29, no. 4, pp. 1–30, Oct. 2021.
- [97] S. Dash, O. Gunluk, and D. Wei, "Boolean decision rules via column generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
- [98] D. Wei, S. Dash, T. Gao, and O. Gunluk, "Generalized linear rule models," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6687–6696.
- [99] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 260–269.
- [100] H. Li, F. Wei, and H. Hu, "Enabling dynamic network access control with anomaly-based IDS and SDN," in *Proc. ACM Int. Workshop Secur. Softw. Defined Netw. Netw. Function Virtualization (SDN-NFVSec)*, 2019, pp. 13–16.
- [101] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "LEMNA: Explaining deep learning based security applications," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 364–379.
- [102] K. Amarasinghe and M. Manic, "Improving user trust on deep neural networks based intrusion detection systems," in *Proc. 44th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2018, pp. 3262–3268.
- [103] T. Zebin, S. Rezvy, and Y. Luo, "An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2339–2349, 2022.
- [104] A. Morichetta, P. Casas, and M. Mellia, "EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis," in *Proc. 3rd ACM CoNEXT Workshop Big Data, Mach. Learn. Artif. Intell. Data Commun. Netw.*, Dec. 2019, pp. 22–28.
- [105] G. Andresini, A. Appice, F. P. Caforio, D. Malerba, and G. Vessio, "ROULETTE: A neural attention multi-output model for explainable network intrusion detection," *Exp. Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117144.
- [106] A. A. Reyes, F. D. Vaca, G. A. Castro Aguayo, Q. Niyaz, and V. Devabhaktuni, "A machine learning based two-stage Wi-Fi network intrusion detection system," *Electronics*, vol. 9, no. 10, p. 1689, Oct. 2020.
- [107] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," 2021, *arXiv:2103.07110*.
- [108] M. Sarhan, S. Layeghy, and M. Portmann, "Evaluating standard feature sets towards increased generalisability and explainability of ML-based network intrusion detection," 2021, *arXiv:2104.07183*.
- [109] N. I. Mowla, J. Rosell, and A. Vahidi, "Dynamic voting based explainable intrusion detection system for in-vehicle network," in *Proc. 24th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2022, pp. 406–411.
- [110] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin, "TRUST XAI: Model-agnostic explanations for AI with a case study on IIoT security," *IEEE Internet Things J.*, early access, Oct. 21, 2022, doi: [10.1109/JIOT.2021.3122019](https://doi.org/10.1109/JIOT.2021.3122019).
- [111] B. Mahbooba, R. Sahal, W. Alosaimi, and M. Serrano, "Trust in intrusion detection systems: An investigation of performance analysis for machine learning and deep learning models," *Complexity*, vol. 2021, pp. 1–23, Mar. 2021.
- [112] M. Rabbani, Y. L. Wang, R. Khoshkangini, H. Jelodar, R. Zhao, and P. Hu, "A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing," *J. Netw. Comput. Appl.*, vol. 151, Feb. 2020, Art. no. 102507.
- [113] D. Arivudainambi, V. K. Ka, and P. Visu, "Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance," *Comput. Commun.*, vol. 147, pp. 50–57, Nov. 2019.
- [114] A. Namavar Jahromi, S. Hashemi, A. Dehghantaha, K.-K.-R. Choo, H. Karimpour, D. E. Newton, and R. M. Parizi, "An improved two-hidden-layer extreme learning machine for malware hunting," *Comput. Secur.*, vol. 89, Feb. 2020, Art. no. 101655.
- [115] M. Alaeiyan, S. Parsa, and M. Conti, "Analysis and classification of context-based malware behavior," *Comput. Commun.*, vol. 136, pp. 76–90, Feb. 2019.
- [116] J. Stiborek, T. Pevný, and M. Reháč, "Multiple instance learning for malware classification," *Exp. Syst. Appl.*, vol. 93, pp. 346–357, Mar. 2018.

- [117] L. Xiaofeng, J. Fangshuo, Z. Xiao, Y. Shengwei, S. Jing, and P. Lio, "ASSCA: API sequence and statistics features combined architecture for malware detection," *Comput. Netw.*, vol. 157, pp. 99–111, Jul. 2019.
- [118] S. Li, Q. Zhou, R. Zhou, and Q. Lv, "Intelligent malware detection based on graph convolutional network," *J. Supercomput.*, vol. 78, no. 3, pp. 4182–4198, Feb. 2022.
- [119] Y. Fan, S. Hou, Y. Zhang, Y. Ye, and M. Abdulhayoglu, "Gotcha-Sly malware!: Scorpion a Metagraph2vec based malware detection system," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 253–262.
- [120] F. Xiao, Z. Lin, Y. Sun, and Y. Ma, "Malware detection based on deep learning of behavior graphs," *Math. Problems Eng.*, vol. 2019, pp. 1–10, Feb. 2019.
- [121] A. G. Kakisim, M. Nar, and I. Sogukpinar, "Metamorphic malware identification using engine-specific patterns based on co-opcode graphs," *Comput. Standards Interfaces*, vol. 71, Aug. 2020, Art. no. 103443.
- [122] R. U. Khan, X. Zhang, and R. Kumar, "Analysis of ResNet and GoogleNet models for malware detection," *J. Comput. Virol. Hacking Techn.*, vol. 15, no. 1, pp. 29–37, 2019.
- [123] D. Nahmias, A. Cohen, N. Nissim, and Y. Elovici, "Deep feature transfer learning for trusted and automated malware signature generation in private cloud environments," *Neural Netw.*, vol. 124, pp. 243–257, Apr. 2020.
- [124] Q. Le, O. Boydel, B. Mac Namee, and M. Scanlon, "Deep learning at the shallow end: Malware classification for non-domain experts," *Digit. Invest.*, vol. 26, pp. S118–S126, Jul. 2018.
- [125] S. Huda, R. Islam, J. Abawajy, J. Yearwood, M. M. Hassan, and G. Fortino, "A hybrid-multi filter-wrapper framework to identify runtime behaviour for fast malware detection," *Future Gener. Comput. Syst.*, vol. 83, pp. 193–207, Jun. 2018.
- [126] I. Baptista, S. Shiaeles, and N. Kolokotronis, "A novel malware detection system based on machine learning and binary visualization," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2019, pp. 1–6.
- [127] F. O. Catak, J. Ahmed, K. Sahinbas, and Z. H. Khand, "Data augmentation based malware detection using convolutional neural networks," *PeerJ Comput. Sci.*, vol. 7, p. e346, Jan. 2021.
- [128] Q. Qian and M. Tang, "Dynamic API call sequence visualisation for malware classification," *IET Inf. Secur.*, vol. 13, no. 4, pp. 367–377, Jul. 2019.
- [129] M. Jain, W. Andreopoulos, and M. Stamp, "Convolutional neural networks and extreme learning machines for malware classification," *J. Comput. Virol. Hacking Techn.*, vol. 16, no. 3, pp. 229–244, Sep. 2020.
- [130] G. Bendiab, S. Shiaeles, A. Alruban, and N. Kolokotronis, "IoT malware network traffic classification using visual representation and deep learning," in *Proc. 6th IEEE Conf. Netw. Softwarization (NetSoft)*, Jun. 2020, pp. 444–449.
- [131] D. Gibert, C. Mateu, J. Planes, and R. Vicens, "Using convolutional neural networks for classification of malware represented as images," *J. Comput. Virol. Hacking Techn.*, vol. 15, no. 1, pp. 15–28, Mar. 2019.
- [132] Y. Ye, L. Chen, S. Hou, W. Hardy, and X. Li, "DeepAM: A heterogeneous deep learning framework for intelligent malware detection," *Knowl. Inf. Syst.*, vol. 54, no. 2, pp. 265–285, Feb. 2018.
- [133] S. Sharma, C. R. Krishna, and S. K. Sahay, "Detection of advanced malware by machine learning techniques," in *Soft Computing: Theories and Applications (Advances in Intelligent Systems and Computing)*, vol. 742, K. Ray, T. Sharma, S. Rawat, R. Saini, and A. Bandyopadhyay, Eds. Singapore: Springer, 2019, doi: 10.1007/978-981-13-0589-4_31.
- [134] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, "DREBIN: Effective and explainable detection of Android malware in your pocket," in *Proc. NDSS*, vol. 14, 2014, pp. 23–26.
- [135] M. Melis, D. Maiorca, B. Biggio, G. Giacinto, and F. Roli, "Explaining black-box Android malware detection," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 524–528.
- [136] M. Kinkead, S. Millar, N. McLaughlin, and P. O'Kane, "Towards explainable CNNs for Android malware detection," *Proc. Comput. Sci.*, vol. 184, pp. 959–965, Jan. 2021.
- [137] R. Kumar, Z. Xiaosong, R. U. Khan, J. Kumar, and I. Ahad, "Effective and explainable detection of Android malware based on machine learning algorithms," in *Proc. Int. Conf. Comput. Artif. Intell. (ICCAI)*, 2018, pp. 35–40.
- [138] B. Wu, S. Chen, C. Gao, L. Fan, Y. Liu, W. Wen, and M. R. Lyu, "Why an Android APP is classified as malware: Toward malware classification interpretation," *ACM Trans. Softw. Eng. Methodol.*, vol. 30, no. 2, pp. 1–29, Apr. 2021.
- [139] D. Zhu, T. Xi, P. Jing, D. Wu, Q. Xia, and Y. Zhang, "A transparent and multimodal malware detection method for Android apps," in *Proc. 22nd Int. ACM Conf. Model., Anal. Simul. Wireless Mobile Syst. (MSWIM)*, 2019, pp. 51–60.
- [140] J. Feichtner and S. Gruber, "Understanding privacy awareness in Android APP descriptions using deep learning," in *Proc. 10th ACM Conf. Data Appl. Secur. Privacy*, 2020, pp. 203–214.
- [141] G. Iadarola, F. Martinelli, F. Mercaldo, and A. Santone, "Towards an interpretable deep learning model for mobile malware detection and family identification," *Comput. Secur.*, vol. 105, Jun. 2021, Art. no. 102198.
- [142] S. Chen, S. Bateni, S. Grandhi, X. Li, C. Liu, and W. Yang, "DENAS: Automated rule generation by knowledge extraction from neural networks," in *Proc. 28th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, Nov. 2020, pp. 813–825.
- [143] L. Yang, W. Guo, Q. Hao, A. Ciptadi, A. Ahmadzadeh, X. Xing, and G. Wang, "CADE: Detecting and explaining concept drift samples for security applications," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 2327–2344.
- [144] Z. Pan, J. Sheldon, and P. Mishra, "Hardware-assisted malware detection using explainable machine learning," in *Proc. IEEE 38th Int. Conf. Comput. Design (ICCD)*, Oct. 2020, pp. 663–666.
- [145] Z. Pan, J. Sheldon, and P. Mishra, "Hardware-assisted malware detection and localization using explainable machine learning," *IEEE Trans. Comput.*, early access, Feb. 11, 2022, doi: 10.1109/TC.2022.3150573.
- [146] S. Bose, T. Barao, and X. Liu, "Explaining AI for malware detection: Analysis of mechanisms of MalConv," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.
- [147] M. Al-Faw'rah, A. Saif, M. T. Jafar, and A. Elhassan, "Malware detection by eating a whole APK," in *Proc. 32nd Int. Conf. for Internet Technol. Secured Trans. (ICITST)*, Dec. 2020, pp. 1–7.
- [148] B. Hsupeng, K.-W. Lee, T.-E. Wei, and S.-H. Wang, "Explainable malware detection using predefined network flow," in *Proc. 24th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2022, pp. 27–33.
- [149] W. Han, J. Xue, Y. Wang, L. Huang, Z. Kong, and L. Mao, "MalDAE: Detecting and explaining malware based on correlation and fusion of static and dynamic characteristics," *Comput. Secur.*, vol. 83, pp. 208–233, Jun. 2019.
- [150] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, and A. Armando, "Explaining vulnerabilities of deep learning to adversarial malware binaries," 2019, *arXiv:1901.03583*.
- [151] I. Rosenberg, S. Meir, J. Berrebi, I. Gordon, G. Sicard, and E. O. David, "Generating end-to-end adversarial examples for malware classifiers using explainability," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–10.
- [152] G. Severi, J. Meyer, S. Coull, and A. Oprea, "Explanation-Guided backdoor poisoning attacks against malware classifiers," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 1487–1504.
- [153] W. Song, X. Li, S. Afroz, D. Garg, D. Kuznetsov, and H. Yin, "Automatic generation of adversarial examples for interpreting malware classifiers," 2020, *arXiv:2003.03100*.
- [154] M. Fan, W. Wei, X. Xie, Y. Liu, X. Guan, and T. Liu, "Can we trust your explanations? Sanity checks for interpreters in Android malware analysis," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 838–853, 2021.
- [155] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, May 2019.
- [156] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Exp. Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019.
- [157] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," in *Proc. 3rd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Mar. 2020, pp. 1–6.
- [158] P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web phishing detection using a deep learning framework," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–9, Sep. 2018.
- [159] E. Zhu, Y. Ju, Z. Chen, F. Liu, and X. Fang, "DFOB-ANN: An artificial neural network phishing detection model based on decision tree and optimal features," *Appl. Soft Comput.*, vol. 95, Oct. 2020, Art. no. 106505.

- [160] R. S. Rao and A. R. Pais, "Jail-phish: An improved search engine based phishing detection system," *Comput. Secur.*, vol. 83, pp. 246–267, Jun. 2019.
- [161] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020.
- [162] H. Faris, H. Faris, A.-Z. Ala'M, A. A. Heidari, I. Aljarah, M. Mafarja, M. A. Hassonah, and H. Fujita, "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks," *Inf. Fusion*, vol. 48, pp. 67–83, Aug. 2019.
- [163] G. Chetty, H. Bui, and M. White, "Deep learning based spam detection system," in *Proc. Int. Conf. Mach. Learn. Data Eng. (iCMLDE)*, Dec. 2019, pp. 91–96.
- [164] A. Barushka and P. Hajek, "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks," *Appl. Intell.*, vol. 48, no. 10, pp. 3538–3556, Oct. 2018.
- [165] S. Douzi, F. A. AlShahwan, M. Lemoudden, and B. Ouahidi, "Hybrid email spam detection model using artificial intelligence," *Int. J. Mach. Learn. Comput.*, vol. 10, no. 2, pp. 316–322, Feb. 2020.
- [166] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, Jan. 2019.
- [167] S. Magdy, Y. Abouelseoud, and M. Mikhail, "Efficient spam and phishing emails filtering based on deep learning," *Comput. Netw.*, vol. 206, Apr. 2022, Art. no. 108826.
- [168] S. Bosaeed, I. Katib, and R. Mehmood, "A fog-augmented machine learning based SMS spam detection and classification system," in *Proc. 5th Int. Conf. Fog Mobile Edge Comput. (FMEC)*, Apr. 2020, pp. 325–330.
- [169] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 3793–3810.
- [170] S. Mahdaviifar and A. A. Ghorbani, "Dennes: Deep embedded neural network expert system for detecting cyber attacks," *Neural Comput. Appl.*, vol. 32, no. 18, pp. 14753–14780, 2020.
- [171] Y. Chai, Y. Zhou, W. Li, and Y. Jiang, "An explainable multi-modal hierarchical attention model for developing phishing threat intelligence," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 790–803, Apr. 2022.
- [172] K. Kluge and R. Eckhardt, "Explaining the suspicion: Design of an XAI-based user-focused anti-phishing measure," in *Innovation Through Information Systems. WI 2021 (Lecture Notes in Information Systems and Organisation)*, vol. 47, F. Ahlemann, R. Schütte, and S. Stieglitz, Eds. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-86797-3_17.
- [173] P. R. G. Fernandes, C. P. Floret, K. F. C. De Almeida, V. C. Da Silva, J. P. Papa, and K. A. P. Da Costa, "Phishing detection using URL-based XAI techniques," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2021, pp. 01–06.
- [174] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A unified framework for machine learning interpretability," 2019, *arXiv:1909.09223*.
- [175] M. Stites, M. Nyre-Yu, B. Moss, C. Smutz, and M. Smith, "Sage advice? The impacts of explanations for machine learning models on human decision-making in spam detection," in *Proc. Int. Conf. Hum.-Comput. Interact.*, Jul. 2021, pp. 269–284.
- [176] D. Zhang, Q. Zhang, G. Zhang, and J. Lu, "FreshGraph: A spam-aware recommender system for cold start problem," in *Proc. IEEE 14th Int. Conf. Intell. Syst. Knowl. Eng. (ISKE)*, Nov. 2019, pp. 1211–1218.
- [177] J. Gu, J. Na, J. Park, and H. Kim, "Predicting success of outbound telemarketing in insurance policy loans using an explainable multiple-filter convolutional neural network," *Appl. Sci.*, vol. 11, no. 15, p. 7147, Aug. 2021.
- [178] T. Le, S. Wang, and D. Lee, "GRACE: Generating concise and informative contrastive sample to explain neural network model's prediction," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 238–248.
- [179] A. Björklund, J. Mäkelä, and K. Puolamäki, "SLISEMAP: Supervised dimensionality reduction through local explanations," 2022, *arXiv:2201.04455*.
- [180] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Exp. Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117193.
- [181] A. Capillo, E. de Santis, F. Mascioli, and A. Rizzi, "Mining M-grams by a granular computing approach for text classification," in *Proc. 12th Int. Joint Conf. Comput. Intell.*, 2020, pp. 350–360.
- [182] S. Cresci, "A decade of social bot detection," *Commun. ACM*, vol. 63, no. 10, pp. 72–83, Sep. 2020.
- [183] H. Owen, J. Zarrin, and S. M. Pour, "A survey on botnets, issues, threats, methods, detection and prevention," *J. Cybersecurity Privacy*, vol. 2, no. 1, pp. 74–88, Feb. 2022.
- [184] A. Almomani, "Fast-flux hunter: A system for filtering online fast-flux botnet," *Neural Comput. Appl.*, vol. 29, no. 7, pp. 483–493, Apr. 2018.
- [185] X. Pei, S. Tian, L. Yu, H. Wang, and Y. Peng, "A two-stream network based on capsule networks and sliced recurrent neural networks for DGA botnet detection," *J. Netw. Syst. Manage.*, vol. 28, no. 4, pp. 1694–1721, Oct. 2020.
- [186] S. I. Popoola, B. Adebisi, R. Ande, M. Hammoudeh, and A. A. Atayero, "Memory-efficient deep learning for botnet attack detection in IoT networks," *Electronics*, vol. 10, no. 9, p. 1104, May 2021.
- [187] V. A. Memos and K. E. Psannis, "AI-powered honeypots for enhanced IoT botnet detection," in *Proc. 3rd World Symp. Commun. Eng. (WSCSE)*, Oct. 2020, pp. 64–68.
- [188] W. Jung, H. Zhao, M. Sun, and G. Zhou, "IoT botnet detection via power consumption modeling," *Smart Health*, vol. 15, Mar. 2020, Art. no. 100103.
- [189] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "RTbust: Exploiting temporal patterns for botnet detection on Twitter," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 183–192.
- [190] C. Joshi, R. Ranjan, and V. Bharti, "A fuzzy logic based feature engineering approach for botnet detection using ANN," *J. King Saud Univ. Comput. Inf. Sci.*, pp. 1–11, Jul. 2021, doi: 10.1016/j.jksuci.2021.06.018.
- [191] H.-T. Nguyen, Q.-D. Ngo, D.-H. Nguyen, and V.-H. Le, "PSI-rooted subgraph: A novel feature for IoT botnet detection using classifier algorithms," *ICT Exp.*, vol. 6, no. 2, pp. 128–138, Jun. 2020.
- [192] M. M. Alani, "BotStop : Packet-based efficient and explainable IoT botnet detection using machine learning," *Comput. Commun.*, vol. 193, pp. 53–62, Sep. 2022.
- [193] P. P. Kundu, T. Truong-Huu, L. Chen, L. Zhou, and S. G. Teo, "Detection and classification of botnet traffic using deep learning with model explanation," *IEEE Trans. Dependable Secure Comput.*, early access, Jun. 15, 2022, doi: 10.1109/TDSC.2022.3183361.
- [194] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, "Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intelligence sharing," *IEEE Access*, vol. 10, pp. 34613–34624, 2022.
- [195] N. Ben Rabah, B. Le Grand, and M. K. Pinheiro, "IoT botnet detection using black-box machine learning models: The trade-off between performance and interpretability," in *Proc. IEEE 30th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises (WETICE)*, Oct. 2021, pp. 101–106.
- [196] A. Guerra-Manzanares, S. Nomm, and H. Bahsi, "Towards the integration of a post-hoc interpretation step into the machine learning workflow for IoT botnet detection," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1162–1169.
- [197] X. Zhu, Y. Zhang, Z. Zhang, D. Guo, Q. Li, and Z. Li, "Interpretability evaluation of botnet detection model based on graph neural network," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2022, pp. 1–6.
- [198] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19620–19631.
- [199] M. Zago, M. G. Pérez, and G. M. Pérez, "Early DGA-based botnet identification: Pushing detection to the edges," *Cluster Comput.*, vol. 24, no. 3, pp. 1695–1710, Sep. 2021.
- [200] A. Drichel, N. Faerber, and U. Meyer, "First step towards EXPLAINable DGA multiclass classification," in *Proc. 16th Int. Conf. Availability, Rel. Secur.*, Aug. 2021, pp. 1–13.
- [201] F. Becker, A. Drichel, C. Müller, and T. Ertl, "Interpretable visualizations of deep neural networks for domain generation algorithm detection," in *Proc. IEEE Symp. Visualizat. Cyber Secur. (VizSec)*, Oct. 2020, pp. 25–29.
- [202] Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "GEE: A gradient-based explainable variational autoencoder for network anomaly detection," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2019, pp. 91–99.

- [203] M. Kouvela, I. Dimitriadis, and A. Vakali, "Bot-detective: An explainable Twitter bot detection service with crowdsourcing functionalities," in *Proc. 12th Int. Conf. Manage. Digit. EcoSystems*, Nov. 2020, pp. 55–63.
- [204] C. Khanan, W. Luwichana, K. Pruktharithikoon, J. Jiarpakdee, C. Tantithamthavorn, M. Choetkiertikul, C. Ragkhitwetsagul, and T. Sunetnanta, "JITBOT: An explainable just-in-time defect prediction bot," in *Proc. 35th IEEE/ACM Int. Conf. Automated Softw. Eng.*, Sep. 2020, pp. 1336–1339.
- [205] I. Dimitriadis, K. Georgiou, and A. Vakali, "Social botomics: A systematic ensemble ML approach for explainable and multi-class bot detection," *Appl. Sci.*, vol. 11, no. 21, p. 9857, Oct. 2021.
- [206] E. Park, K. Ho Park, and H. Kang Kim, "Understand watchdogs: Discover how game bot get discovered," 2020, *arXiv:2011.13374*.
- [207] D. B. Lira, F. Xavier, and L. A. Digiampietri, "Combining clustering and classification algorithms for automatic bot detection: A case study on posts about COVID-19," in *Proc. 17th Brazilian Symp. Inf. Syst.*, Jun. 2021, pp. 1–7.
- [208] S. X. Rao, S. Zhang, Z. Han, Z. Zhang, W. Min, Z. Chen, Y. Shan, Y. Zhao, and C. Zhang, "xFraud: Explainable fraud transaction detection," *Proc. VLDB Endowment*, no. 3, pp. 427–436, Nov. 2021.
- [209] T. Srinath and H. Gururaja, "Explainable machine learning in identifying credit card defaulters," *Global Transitions Proc.*, vol. 3, no. 1, pp. 119–126, Jun. 2022.
- [210] P. Biecek, "DALEX: Explainers for complex predictive models in R," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 3245–3249, 2018.
- [211] S. Venkatraman and M. Alazab, "Use of data visualisation for zero-day malware detection," *Secur. Commun. Netw.*, vol. 2018, pp. 1–13, Dec. 2018.
- [212] R. Kumar and G. Subbiah, "Zero-day malware detection and effective malware analysis using Shapley ensemble boosting and bagging approach," *Sensors*, vol. 22, no. 7, p. 2798, Apr. 2022.
- [213] J. H. Sejr, A. Zimek, and P. Schneider-Kamp, "Explainable detection of zero day web attacks," in *Proc. 3rd Int. Conf. Data Intell. Secur. (ICDIS)*, Jun. 2020, pp. 71–78.
- [214] Q. Zhou, R. Li, L. Xu, A. Nallanathan, J. Yang, and A. Fu, "Towards explainable meta-learning for DDoS detection," 2022, *arXiv:2204.02255*.
- [215] S. W. Hall, A. Sakzad, and K. R. Choo, "Explainable artificial intelligence for digital forensics," *WIREs Forensic Sci.*, vol. 4, no. 2, Mar. 2022.
- [216] Y. S. Pethe and P. R. Dandekar, "ATLE2FC: Design of an augmented transfer learning model for explainable IoT forensics using ensemble classification," in *Proc. Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC)*, May 2022, pp. 131–137.
- [217] C. Kraetzer, D. Siegel, S. Seidlitz, and J. Dittmann, "Process-driven modelling of media forensic investigations—considerations on the example of DeepFake detection," *Sensors*, vol. 22, no. 9, p. 3137, Apr. 2022.
- [218] S. Dennis, K. Christian, S. Stefan, and D. Jana, "Forensic data model for artificial intelligence based media forensics—Illustrated on the example of DeepFake detection," *Electron. Imag.*, vol. 34, pp. 1–6, Jan. 2022.
- [219] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, "Explainable unsupervised machine learning for cyber-physical systems," *IEEE Access*, vol. 9, pp. 131824–131843, 2021.
- [220] P. R. Aryan, F. J. Ekaputra, M. Sabou, D. Hauer, R. Mosshammer, A. Einfalt, T. Miksa, and A. Rauber, "Explainable cyber-physical energy systems based on knowledge graph," in *Proc. 9th Workshop Model. Simul. Cyber-Phys. Energy Syst.*, May 2021, pp. 1–6.
- [221] M. Blumreiter, J. Greenyer, F. J. C. Garcia, V. Klos, M. Schwammler, C. Sommer, A. Vogelsang, and A. Wortmann, "Towards self-explainable cyber-physical systems," in *Proc. ACM/IEEE 22nd Int. Conf. Model Driven Eng. Lang. Syst. Companion (MODELS-C)*, Sep. 2019, pp. 543–548.
- [222] R. R. Karn, P. Kudva, H. Huang, S. Suneja, and I. M. Elfadel, "Cryptomining detection in container clouds using system calls and explainable machine learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 3, pp. 674–691, Mar. 2021.
- [223] M. Tavallae, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.
- [224] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 446–452, 2015.
- [225] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguous and discontinuous system call patterns," *IEEE Trans. Comput.*, vol. 63, no. 4, pp. 807–819, Apr. 2014.
- [226] G. Creech, "Developing a high-accuracy cross platform host-based intrusion detection system capable of reliably detecting zero-day attacks," Ph.D. dissertation, School Eng. Inf. Technol., Univ. College, Univ. New South Wales, Austral. Defence Force Acad., Sydney, NSW, Australia, 2014. [Online]. Available: <http://handle.unsw.edu.au/1959.4/53218>
- [227] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–6.
- [228] C. Koliass, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 184–208, 1st. Quart., 2016.
- [229] R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems," *Int. J. Eng. Technol.*, vol. 7, pp. 479–482, Dec. 2018.
- [230] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy*, vol. 1, Jan. 2018, pp. 108–116.
- [231] T. M. Kebede, O. Djaneye-Boundjou, B. N. Narayanan, A. Ralescu, and D. Kapp, "Classification of malware programs using autoencoders based deep learning architecture and its application to the Microsoft malware classification challenge (BIG 2015) dataset," in *Proc. IEEE Nat. Aerosp. Electron. Conf. (NAECON)*, Jun. 2017, pp. 70–75.
- [232] H. S. Anderson and P. Roth, "EMBER: An open dataset for training static PE malware machine learning models," 2018, *arXiv:1804.04637*.
- [233] G. Severi, T. Leek, and B. Dolan-Gavitt, "MALREC: Compact full-trace malware recording for retrospective deep analysis," in *Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2018 (Lecture Notes in Computer Science)*, vol. 10885, C. Giuffrida, S. Bardin, and G. Blanc, Eds. Cham, Switzerland: Springer, 2018, doi: [10.1007/978-3-319-93411-2_1](https://doi.org/10.1007/978-3-319-93411-2_1).
- [234] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi, "Microsoft malware classification challenge," 2018, *arXiv:1802.10135*.
- [235] L. Taheri, A. F. A. Kadir, and A. H. Lashkari, "Extensible Android malware detection and family classification using network-flows and API-calls," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2019, pp. 1–8.
- [236] G. Sakkis, I. Androutopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "A memory-based approach to anti-spam filtering for mailing lists," *Inf. Retr.*, vol. 6, no. 1, pp. 49–73, 2003.
- [237] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004 (Lecture Notes in Computer Science)*, vol. 3201, J. F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Berlin, Germany: Springer, 2004, doi: [10.1007/978-3-540-30115-8_22](https://doi.org/10.1007/978-3-540-30115-8_22).
- [238] R. Shams and R. E. Mercer, "Classifying spam emails using text and readability features," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 657–666.
- [239] D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, A. Ghorbani, and D. Garant, "Botnet detection based on traffic behavior analysis and flow intervals," *Comput. Secur.*, vol. 39, pp. 2–16, Nov. 2013.
- [240] M. Zago, M. G. Pérez, and G. M. Pérez, "UMUDGA: A dataset for profiling algorithmically generated domain names in botnet detection," *Data Brief*, vol. 30, Jun. 2020, Art. no. 105400.
- [241] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, 2014, pp. 1–10.
- [242] E. Holder and N. Wang, "Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst," *Hum.-Intell. Syst. Integr.*, vol. 3, no. 2, pp. 139–153, Jun. 2021.
- [243] A. Kuppa and N.-A. Le-Khac, "Adversarial XAI methods in cybersecurity," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4924–4938, 2021.
- [244] J. Vadillo, R. Santana, and J. A. Lozano, "When and how to fool explainable models (and humans) with adversarial examples," 2021, *arXiv:2107.01943*.



NICOLA CAPUANO received the degree in computer science and the Ph.D. degree in computer science and information engineering. He is currently an Assistant Professor at the School of Engineering, University of Basilicata, Italy. He is the author of about 120 publications in scientific journals, conference proceedings, and books. His research interests include computational intelligence, AI in education, knowledge-based systems, and cognitive robotics. He is an Executive Committee Member

of the Learning Ideas Conference, as well as a scientific referee and a member of the editorial board for several other international journals and conferences. He is an Associate Editor of the *Journal of Ambient Intelligence and Humanized Computing* and *Frontiers in Artificial Intelligence*.



GIUSEPPE FENZA (Member, IEEE) received the degree and Ph.D. degrees in computer sciences from the University of Salerno, Italy, in 2004 and 2009, respectively. He is currently an Associate Professor of computer science at the University of Salerno. The research activity concerns computational intelligence methods to support semantic-enabled solutions and decision-making. He has over 60 publications in fuzzy decision making, knowledge extraction and management, situa-

tion and context awareness, semantic information retrieval, service oriented architecture, and ontology learning. More recently, he worked in automating open source intelligence and big data analytics for counterfeiting extremism and supporting information disorder awareness.



VINCENZO LOIA (Senior Member, IEEE) received the degree in computer science from the University of Salerno, Italy, in 1985, and the Ph.D. degree in computer science from the Université Pierre & Marie Curie Paris VI, France, in 1989. He is currently a Computer Science Full Professor at the University of Salerno, where he worked as a Researcher, from 1989 to 2000, and as an Associate Professor, from 2000 to 2004. He is the Co-Editor-in-Chief of *Soft Computing* and the Editor-in-Chief of *Journal of Ambient Intelligence and Humanized Computing*. He serves as an editor for 14 other international journals.



CLAUDIO STANZIONE (Member, IEEE) received the bachelor's degree in economics and business management and the master's degree in economics from the University of Salerno, Italy, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree in innovation sciences for defence and security—digital transformation and cybersecurity with the Center for Higher Defence Studies (CASD). His research interests include explainable artificial intelligence, with a view in

cyber security applications to analyze the existing methods and literature in order to achieve a greater transparency in military and cyber security fields.

...