Explainable Cyber Defences in the Internet of Things Networks: Opportunities and Solutions

Nour Moustafa, *Senior, IEEE*, Nam Pham, Izhar Ahmed Khan, and Albert Y. Zomaya, *Fellow, IEEE*

***Abstract*—Although the field of explainable artificial intelligence (XAI) has a great interest these days, XAI for cyber security applications still needs further investigation to understand how attack surfaces and vectors would be discovered. In cyber defences, especially anomaly-based intrusion detection systems (IDS), the emerging applications of machine/deep learning models require the interpretation of the models' architecture and the explanation of models' prediction to examine how cyberattacks would occur. This paper presents a comprehensive review of XAI techniques for anomaly-based intrusion detection in the Internet of Things (IoT) networks. Firstly, we review intrusion detection systems with a focus on anomaly-based detection techniques in IoT networks, and how these techniques can be utilised by XAI models. Secondly, we review AI, including machine learning (ML) and deep learning (DL), for anomaly detection applications and IoT ecosystems. Moreover, we discuss how DL's ability to effectively learn from large-scale IoT network datasets, accomplishing high performances in the discovery and interpretation of security events. Thirdly, we demonstrate recent research studies in the intersection of XAI, anomaly-based IDS and IoT networks. Finally, we discuss current challenges and solutions of XAI for security applications, revealing future research directions and illustrating that new security applications demand XAI models to assist decision-makers in understanding and explaining security events.**

***Index Terms*—Intrusion Detection System (IDS), Artificial Intelligence (AI), Explainable AI (XAI), Internet of Things (IoT), Cyber defence**

I. Introduction

Nowadays, due to the Fourth Industrial Revolution, communication technologies and digital devices have been available and affordable in various industries, they has became very popular and important to human's normal activities [1]. As human life is becoming more dependent on digital appliances, the security and privacy of information technology systems and networks are crucial to any organisation [2]. Internet of Things (IoT) networks, including sensors and actuators linked to the Internet, would improve human life quality, but they contain a wide range of vulnerabilities for malicious purposes [3]. IoT security issues are more challenging than traditional ones because of its ubiquity deployment as a distributed network [4]. Moreover, due to the proliferation of heterogeneous devices, IoT networks generate high-dimensional and multimodal data, which requires the capability of analysing big data. Artificial Intelligence (AI), especially Machine Learning (ML) and Deep Learning (DL), can attain this requirement as they have been utilised across industries and achieved excellent performances with large scales of data [5].

Due to the potential benefits gained from compromising computer systems, malicious actors invest a lot of money, time and effort in making sophisticated cyber attacks. Zero-day attacks and multi-stage attacks are significant challenges to ensuring digital assets' security as they can bypass traditional security mechanisms. While zero-day attacks utilise unknown malicious techniques, multi-stage attacks combine multiple phases in which each step taken is insufficient to be recognised as malicious [6]. Advanced Persistent Threats (APTs) would use even more complex techniques to achieve their malicious intent [7].

Each day, millions of IoT devices around the world exchanges data, creating large-scale communication and potential targets for intruders [8]. As a results, cyber defence is crucial to protect data and information systems against cybercrime. Cyber defence is the process of developing threat hunting that can be managed using intrusion detection system (IDS). Threat hunting refers to the proactive search for cyber threats in a particular environment based on expertise knowledge and forensics [9]. IoT networks would include many vulnerabilities that make them challenging to implement cyber defence mechanisms due to the heterogeneous and resource-constraint nature of the IoT ecosystem.

Multiple security mechanisms have been proposed [10] to protect IoT network. Traditional mechanisms, such as anti-malware, firewalls, authentication and encryption, are all well-known, where each of them fits different purposes [10]. However, these mechanisms are ineffective, due to lacking dynamism and the rapid growth of attack techniques [9]. Moreover, it is insufficient to deploy these conventional methods for mega systems that combine many devices with inherent vulnerabilities, such as IoT environments [11]. An Intrusion Detection System (IDS) is a prominent approach that can identify diverse cyber attacks and even zero-day attacks. As a result, recent years have witnessed many advancements in IDSs. Due to the application of AI techniques, AI-based IDSs can achieve good performance with benchmark datasets. Deep learning techniques can analyse complex data and learn from previous attack patterns in the dataset to detect zero-day attacks. However, such techniques suffer from a high false-positive rate and opaqueness to users.

Anomaly-based detection is important to prevent unknown attacks and protect information technology systems. Therefore,

Nour Moustafa and Nam Pham are with the School of Engineering and Information Technology, University of New South Wales, Canberra, Australia. E-mail:thenam.pham@student.unsw.edu.au, nour.moustafa@unsw.edu.au.

Izhar Ahmed Khan is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, People's Republic of China. E-mail:izhar@nuaa.edu.cn.

Albert Y. Zomaya is with Centre for Distributed and High Performance Computing, School of Information Technologies, University of Sydney, Sydney, NSW, Australia. Email: albert.zomaya@sydney.edu.au.

many ML algorithms have been proposed to design efficient IDSs with high accuracy and low false-positive rates. Moreover, the need to explain the functioning and predictions of ML-based IDS arises as different types of users are benefited from understanding the root cause of intrusion detection. Therefore, the field of explainable artificial intelligence (XAI) has been developed in recent years to address this concern. XAI illuminates black-box models of ML by providing explanations on their functioning and predictions [12], [13].

Research Motivation– The challenge of developing XAI for anomaly-based IDSs is a combination of difficulties in developing an effective anomaly-based IDS and obtaining explainability by AI techniques. On the one hand, the challenges for designing effective IDSs mainly consist of a comprehensive dataset and real-time detection. The dataset is crucial in anomaly-based IDS as it is used to train the ML model and significantly affects the model's performance in real-life deployment; however, collecting a comprehensive and high-quality dataset that reflects all possible types of intrusion is impossible [3]. Moreover, real-time detection capability is challenging to be achieved as it causes a long processing time and high false alarm rate [14]. On the other hand, designing an efficient XAI method that can be easily accessible and evaluated is difficult. XAI needs to generate explanations for multiple types of the recipient so that the model's predictions can be thoroughly understood; however, existing literature only focuses on a few types of audience such as researchers and developers [15]. Moreover, the lack of standardised terminologies and definitions in the field causes difficulties in the evaluation of the models' performances.

Research Contribution–This paper presents a comprehensive survey that discusses XAI methods and techniques for developing explainable cyber defences, especially anomaly-based IDS, in IoT networks. The major contributions of this study include the follows:

1) We demonstrate intrusion detection systems with a focus on anomaly-based detection in IoT networks. Different types of IDS are categorised and analysed. Their utilisation of these systems for various IoT environments are discussed, demonstrating their advantages and disadvantages.
2) We review AI techniques, including machine learning and deep learning, and explain their types. We also explain how these techniques can effectively learn from large-scale datasets to discover and explain cyberattack events
3) The field of XAI and its utilisation for anomaly-based IDS is discussed in-depth. In regards to the need for XAI and its benefits across various sectors, we explore various existing XAI approaches and their potential applications for anomaly-based IDSs in IoT networks.
4) We survey recent studies in the intersection of XAI, anomaly-based IDS and IoT. We show insights into the intersection of those topics and analyse each piece of research.
5) We also identify current challenges and provide future research directions related to XAI for cyber defences in IoT networks. The application of XAI in cybersecurity presents several challenges and provides new opportunities for further research to develop explainable AI-based IDSs in current IoT networks.

This paper consists of seven sections. Section II outlines cybersecurity and terminologies in this field. Section III discusses Intrusion Detection System and Network Anomaly-based Detection System. Section IV discusses Artificial Intelligence and its role in cybersecurity along with Explainable Artificial Intelligence. Section V introduces the Internet of Things and the security challenges to securing IoT networks. Section VI discusses and summarises related work. Lastly, Section VIII concludes the paper.

II. An Overview of Cyber Defences

Over the last decades, there has been an immense increase in the use of computing and digital appliances. People utilize these appliances as they provide convenient and effective means of communication. Ultimately, these devices connect the virtual worlds with the physical worlds, increasing efficiency. Consequently, humans' daily lives heavily depend on these computing networks, applications, or devices [16]. As people become more dependent on the Internet and digital appliances, security issues in this domain become more popular and lead to more severe consequences. According to Cybersecurity Ventures, global cybercrime costs grow by 15% per year and will reach USD 10.5 trillion annually by 2025 [17]. Over the period 1 July 2019 to 30 June 2020, the Australian Cyber Security Centre (ACSC) received 59,806 cybercrime reports at an average of one report every 10 minutes [18]. According to the Australian Competition and Consumer Commissions (ACCC), Australians lost approximately $634 million due to scams in the year 2019 [18]. Microsoft estimated that the cost of cybercrime to the Australian economy is about $29 billion annually, which equivalent to almost 2% of Australia's GDP [19]. This explains why many countries are expected to spend billions of USD for securing the information systems. According to Atlas VPN investigation, the US government is expected to spend $18.78 billion for cybersecurity in 2021 [20].

As mentioned previously, the privacy and security of information systems are of utmost need to any organization. The term cybersecurity is defined in different ways due to the number of different aspects. In [21], the authors define *cybersecurity* as the security and privacy of digital assets, which are everything from computer networks to mobile devices and data that is processed, stored, and transferred by interconnected information systems. Cyber security's goal is to preserve the integrity, confidentiality, and availability of information in cyberspace. In order to comprehensively define cybersecurity, the term cybersecurity refers to the processes, guidelines, technologies, and practices of defending cyberspace from malicious activities. Cyberspace is a global domain in which electronic and electromagnetic spectrum combined with interconnected and dependent networks help create, edit, store and transfer information [22], [23]. The term cyber defence can be declared as the procedure of developing threat hunting and intrusion detection methods and techniques

to protect critical infrastructures of organisations [14]. In this essence, cyber defences should discover and prevent security risks in IoT networks by following general security objectives, including confidentiality, integrity and availability (CIA triad) [24].

Some terminologies in cybersecurity are as follows:

- Vulnerabilities are flaws or weaknesses in a system that attackers can exploit by executing malicious commands, accessing data without having authorization, or conducting denial-of-service attacks [25] [26]. More generally, vulnerabilities refer to any components of the information system that is exploitable and threatens the whole system's security.
- Threats are actions that can be taken from existing vulnerabilities in a system to gain benefit [27]. Different from vulnerabilities, threats would involve outside elements.
- Cyberattack/intrusion is a set of intentional actions taken to exploit an information system using different techniques to compromise confidentiality, integrity, availability of the system and achieves malicious goals [28][29]. Methods to launch a cyber attack can be malware, phishing, SQL injection, Man-In-The-Middle (MITM) and many others.
- Zero-day attacks can be defined as variants of known attacks [30] or unknown attacks that exploit zero-day vulnerabilities in the system. Zero-day vulnerabilities are unknown to the public and are not yet discovered by software vendors or network defenders.
- Multi-stage attack (MSA) is a type of intrusion consisting of a sequence of correlated techniques [6]. The MSA utilises more complex attack techniques over a long period [31] in which each step taken is insufficient to be recognized as aggression. Steps in a typical MSA can be classified into different phases, including reconnaissance, weaponization, delivery, exploitation, installation, command, and control, actions on objective regarding Lockheed Martin's Cyber kill chain [32].
- According to the National Institute of Standards and Technology (NIST), Advanced Persistent Threats (APTs) are adversaries which ultimately aim to steal information, undermine or impede critical aspects of a mission, program, or organisation, or place itself in a position to do so in the future. To do that, APTs establish and extend their presence in information systems by using a wide range of attack surfaces with sophisticated levels of expertise and significant resources. APTs typically launch the attack through multiple steps; they are persistent, targeted attacks on a particular organization [33]. Therefore, APT attacks are a complex version of multi-stage attacks.

III. Intrusion Detection System

Due to the heavy reliance on information systems, various approaches have been deployed to protect information systems such as access control, firewall, anti-malware, sandbox and cryptography [10]. Those are traditional mechanisms that have been widely adopted nowadays; however, due to the rapid growth of attack techniques and zero-day vulnerabilities, such methods are insufficient for securing information systems. Thus, an intrusion detection system (IDS), which is the main method capable of identifying a wide range of cyberattacks, specifically zero-day attacks, has become necessary in any organisation's security infrastructure. *Intrusion detection* is defined as *"the process of monitoring the events occurring in a computer system or network and analysing them for signs of intrusions"* [34]. Bace and Mell (2001) describe intrusions as attempts to compromise the confidentiality, integrity and availability or exploit the security mechanism of digital assets. IDS aims to observe traffic and identify possible threats in the network and computer system by supervising, identifying and evaluating their violations of the security principles [35], [36].

Based on the place of deployment of IDS and the type of system the IDS protects, multiple types of IDS have witnessed a lot of research effort:

- Host-based IDS (HIDS) is attached to the operating system kernel of a specific host and protects the host by forming a layer that allows only legitimate system calls to go through. In other words, HIDS is attached to a single host and watch for malicious activities. In HIDS, computational resources that power the host-based system are taken from the attached host. Host-based methods are reactive, which means that they alert the host after an attack has occurred [37]. Moreover, it might be exposed if the host server is compromised and it cannot be compatible with different platforms [38]. However, HIDS is useful for identifying malicious activities in the organisation's internal pieces of equipment by monitoring system calls, processes, file-system changes and application logs [39]. Moreover, it can analyse the encrypted or obfuscated payloads in the network [40].
- Network-based IDS (NIDS) monitors and inspects activities in the network by reading all inbound packets in the entire network that it is deployed [41]. NIDS do not need to use system resources [37] and can monitor network activities over a particular network segment regardless of the type of the operating system [42], which makes NIDS portable. However, one major drawback of NIDS is that it cannot process encrypted or obfuscated payloads since it only captures information from packet headers [43].
- Cloud-based IDS typically has several different places of deployment. Host-based IDS can be deployed in the virtual machines (VMs) hosted on a cloud server or placed in the hypervisor to monitor the network traffic and the information transferring between the VMs within that hypervisor. Network-based IDS can be deployed to detect abnormalities in the virtual network traffic or to monitor unencrypted network traffic between virtual machines [44].
- IoT-based IDS can be classified into two approaches based on the IDS place of deployment. Centralised IDS is the most widely-used [45] method in which a dedicated central unit such as cloud is utilised to monitor and determine the traffic data in IoT networks. Centralised IoT-based IDS offer the advantages of mighty computation power and centralised management of network traffic;

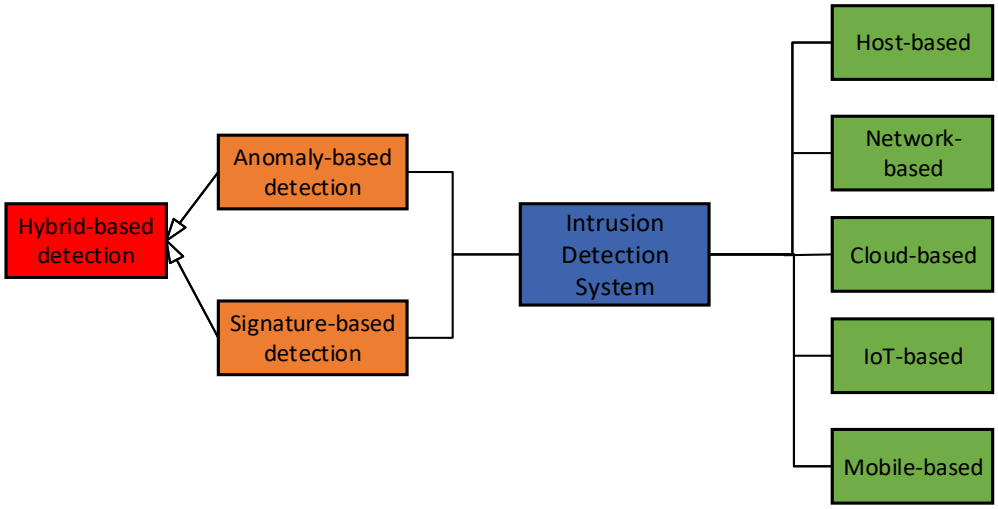

Fig. 1. Classification of IDS

however, it degrades the network performance by generating significant communication overhead [46]. Due to the increasing data traffic that IoT devices generate, this method is gradually becoming unsuitable and replaced. To tackle those problems, decentralised IDS distributes the centralised computing and computational tasks to local fog nodes. Therefore, the heavy load of monitoring traffic data is decreased, and the processing capacity is increased [47].

- Mobile-based IDS has gain popularity in recent years due to the emergence of smartphones. Mobile-based IDS can be further classified into three categories, including (1) host-based: the IDS is deployed on the mobile device, (2) centralised: IDS that is deployed within the cloud will monitor and analyse the mobile devices, (3) distributed: IDS is partly deployed on the cloud and partly on the device [48]. Cloud computing allows centralised data collection and processing; thus, it is convenient and practical to deploy IDS on the cloud to employ the powerful computation power and memory capacity. However, relying on a cloud server has two limitations, including continuous connectivity to the central server and the risk of sensitive information leakage [49].

An IDS can be designed based on three detection methods:

- Signature-based detection recognises possible intrusions by comparing patterns against captured signature of known attacks. Each type of attack usually contains a specific pattern called signature [50]. The signature-based detection method relies on the database consisting of the signature of the existing attack to detect them. As a result, all the existing attacks stored in the database would be detected with high confidence. However, zero-day attacks or even variants of known malicious behaviours would easily bypass the signature-based IDS. Moreover, the database containing the signature of known attacks requires to be updated repeatedly by network security experts; otherwise, this method is not effective [51].
- Anomaly-based detection monitors the everyday activities of network traffic to construct a baseline of unmalicious behaviours. The features or patterns that the anomaly-based IDS model would be static or dynamic can be everything developed by counting the number of packets sent, number of failed attempts to log in, and many others [50]. Whenever any activities have deviated from the constructed normal baseline, an alarm would be generated to alert the admin [52]. Therefore, if the baseline of normal behaviours is built carefully and comprehensively, anomaly-based IDS can detect any types of attacks, including both known and zero-day attacks.
- Hybrid-based detection combines both anomaly-based detection and signature-based detection. Therefore, this method of detection can have advantages from both approaches above. The hybrid-based method depends on the signature-based module to detect known attacks, hence lowering the false alarm rate. At the same time, the anomaly-based module constructs the baseline of normal behaviours in network traffic; thus, it helps detect zero-day attacks.

With the increasing complexity in intrusion techniques, signature-based IDS would be bypassed as attackers utilising zero-day attacks. Anomaly-based IDS can detect zero-day attacks without prior knowledge, which is a significant advantage over the signature-based model. Thus, this paper will focus on network anomaly-based IDSs to build efficient explainable AI-based IDSs, along with high detection accuracy and low false-positive rates.

A. Network Anomaly Detection Systems (NADS)

As mentioned earlier, anomaly-based IDS is effective in detecting known and unknown attacks. It constructs a baseline normal behaviour profile for the monitored network then uses this baseline for comparison of actions at any given time, and the anomalies are reported by raising an alert [3]. A

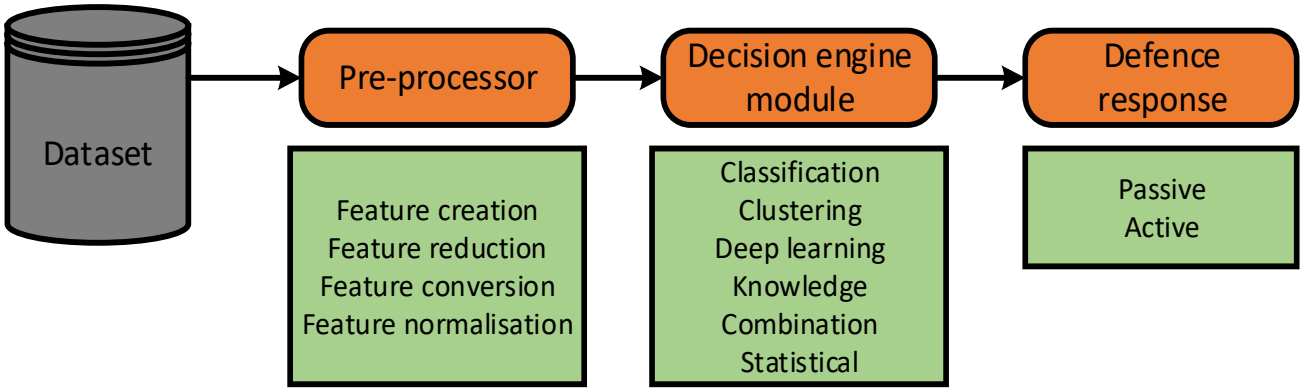

Fig. 2. Components of Network Anomaly Detection

NADS consists of four components: a dataset, data processing module, decision engine (DE) method, and defence responses [53]. A realistic dataset of networks is crucial for building an efficient IDS because it helps to improve the detection accuracy of the IDS in real life and evaluate the performance of the model after training [54].

Building an efficient anomaly-based detection method requires a high-quality data source. However, acquiring a high-quality dataset is challenging due to the difficulty of labelling normal and attack behaviours in live network traffic [55]. Network dataset is collected in a real-time or offline data collection. Then, tools and techniques are employed to store and process a network's big data [56]. For example, to extract network features, several tools are utilised, such as tcpdump, Zeek (previously known as Bro), and MySQL Cluster CGE. The tcpdump can sniff packets on the network, then Zeek extracts the flow-based features from different protocol types in the pcap files. After that, MySQL database is used to store all collected features, and then each record is labelled normal or abnormal [14].

After data collection, there are probably many redundant or duplicated records in the dataset. Therefore, the data processing module is essential in improving the performance of an IDS as it removes noisy and irrelevant information from the collected network data in the dataset. Data-processing consists of the creation, reduction, conversion and normalisation of feature [14].

The DE module is the most critical component of an IDS. In NADS, DE approaches are classified into six categories, including classification-based, clustering-based, deep learning-based, knowledge-based, combination-based, and statistical-based [14]. In all of the mentioned types of DE modules except knowledge-based and statistical-based approaches, there have been attempts to apply machine learning algorithms to distinguish normal and attack events. The performance of the DE module significantly relies on the quality of the data source; however, constructing a comprehensive normal behaviour baseline profile is impossible in real life. Therefore, the anomaly-based detection model suffers from high false-positive rates due to the lack of normal activities profile in the dataset [57].

Defence responses are actions taken after an attack is detected. There are two types of responses, including passive response and active response. In passive response, the network administrator must take action after the IDS raise an alert about malicious behaviour. The form of alarm would be a popup window or an onscreen alert [14]. The active responses refer to a set of actions taken automatically by the IDS that changes the behaviour of an intrusion, such as disconnect users or terminate connections and attacks [58].

IV. Artificial Intelligence for Cyber Security

Artificial Intelligence (AI) is a branch in computer science referring to algorithms that simulate human intelligence in machines capable of imitating human behaviour. In the last decade, the field of AI has seen such rapid growth that AI-based algorithms have been developed in every sector of the technology industry and transformed the way we approach real-world tasks. The advancements in machine learning (ML) and deep learning (DL) are leading this growth.

Especially in the field of cybersecurity, ML/DL techniques offer powerful tools in cybersecurity defence that serve multiple purposes. The use of ML/DL methods in malware or intrusion detection and classification has become common. ML can generalise to never-seen-before malware families, and polymorphic strains [59]. Similarly, it allows anomaly-based IDS to detect zero-day attacks. Tuor et al.[60] use DL techniques for insider threat detection by analysing system logs to detect malicious activities. Additionally, multiple ML methods aid cybersecurity forensic by classifying file fragments [61] and detecting kernel rootkits in Virtual Machines (VMs) [62]. Moreover, Sarker et al.[10] propose a security intelligence modelling using the combination of various AI methods and other techniques. The modelling can be used in multiple domains of cybersecurity to protect against phishing attack and malicious code.

A. Machine Learning

Machine learning (ML) is a subset of AI which builds a mathematical model based on training data in order to make

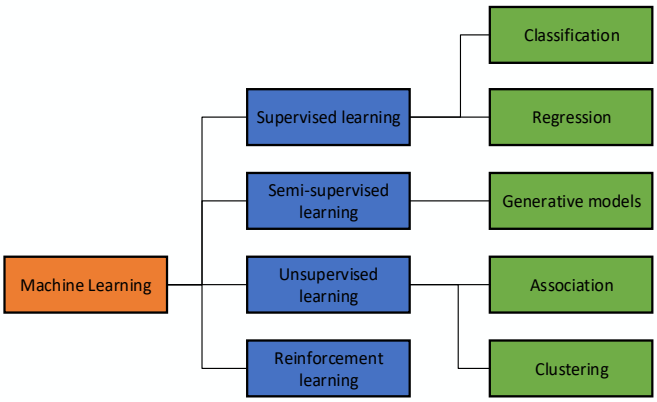

Fig. 3. Categories of Machine Learning

decisions or predictions without being explicitly programmed to perform the task [63]. ML models can perform a wide range of tasks in various fields, including cyber defence. ML techniques have been investigated to optimise various real-world systems, including image recognition, virtual reality, object detection, natural language processing, malware filtering, and many others [64]. Based on the nature of training data or the learning techniques, ML algorithms can be classified into different types as shown in Figure 3.

- **Supervised learning**: the training data has labelled inputs and their desired outputs. After being trained, the ML model can capture the relationships and dependencies between the prediction output and the input features, hence giving correct labels for the features of unknown samples. Supervised learning problems group into classification and regression problems as shown in Figure 3.
- **Unsupervised learning**: the training set only includes inputs without the desired output. The ML model extracts and learns the relationship and patterns in data from unlabeled data on its own. Specifically, the system arranges data into categories or clusters from the offered training figures and input patterns. Therefore, it is also considered as self-organizing, and adaptive learning [65]. There are two main types of unsupervised learning problems, including clustering and association problems.
- **Semi-supervised learning**: this is a combination between supervised learning and unsupervised learning in which the training set includes both unlabeled data and labelled data. Semi-supervised learning is helpful in many scenarios when data collecting is expensive, time-consuming or even unrealistic [66]. This approach can utilize the unlabeled data to improve the learning accuracy [66].
- **Reinforcement learning**: In reinforcement learning, there is no training dataset. The model is trained in a dynamic environment where it interconnects with surrounding by employing trials then receive rewards or penalties depending on its actions. The data feedback is crucial for the model to learn from experience and improve performance. This type of learning is motivated by behaviourist psychology [65].

B. Deep Learning

Deep learning is a subset of machine learning whose architecture is motivated by the structure and functioning of the human brain. Deep learning architecture is multi-layer neural networks. The network consists of multiple layers constructed and connected through neurons. Each neuron is considered a basic computational component, and the whole network presents the computation of the learning process [67]. Several neurons, which usually equals the number of input features, constitute the input layer of the network [68]. The output layer consists of many neurons that equal the number of different classes in the dataset. However, there is usually only one neuron in the binary classification problem, which is a real number in the range 0-1. The last type of layer is hidden layers that are placed between the input layer and output layer. Deep learning models have multiple hidden layers in their network. In [69], the authors classify deep learning models into two categories based on the architectures as follows:

- **Generative**: The unsupervised learning technique is applied to learn from unlabelled data. Generative models depict independence/ dependence for distribution by computing join probability distributions from data with their labels [14]. Models, which can utilize generative architecture, are Recurrent Neural Network (RNN), Deep Auto Encoder (DAE), Deep Boltzmann Machine (DBM) and Deep Belief Network (DBN).
- **Discriminative**: The supervised learning technique is applied to distinguish patterns for prediction tasks. Discriminative models directly estimate the posterior distributions of classes conditioned on the input data [69]. Thus, the discriminative approach is more efficient since it only focuses computational resources on a given task, which is classification, without modelling underlying probability distributions. There are two types of discriminative architecture, including RNN and Convolution Neural Network (CNN).

Generally, DL algorithms demonstrate data as a nested hierarchy of concepts within their architecture of multi-layer neural network. Therefore, deep learning techniques can learn the computational process in depth [14] and achieve good performance and flexibility that outperforms the traditional machine learning in data with high scale [70]. Thus, deep learning approaches are the most suitable for building an efficient IDS.

C. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is an extension of the AI field in which methods are invented to explain ML models' predictions or make models interpretable. The two terms "interpretability" and "explainability" are often used interchangeably by researchers [71]. The field of explainability all started since the publication of paper [12] in the 80s; however, it observed less noticed as AI's growth has focused on predictive performance. In recent years, researchers tend to pay

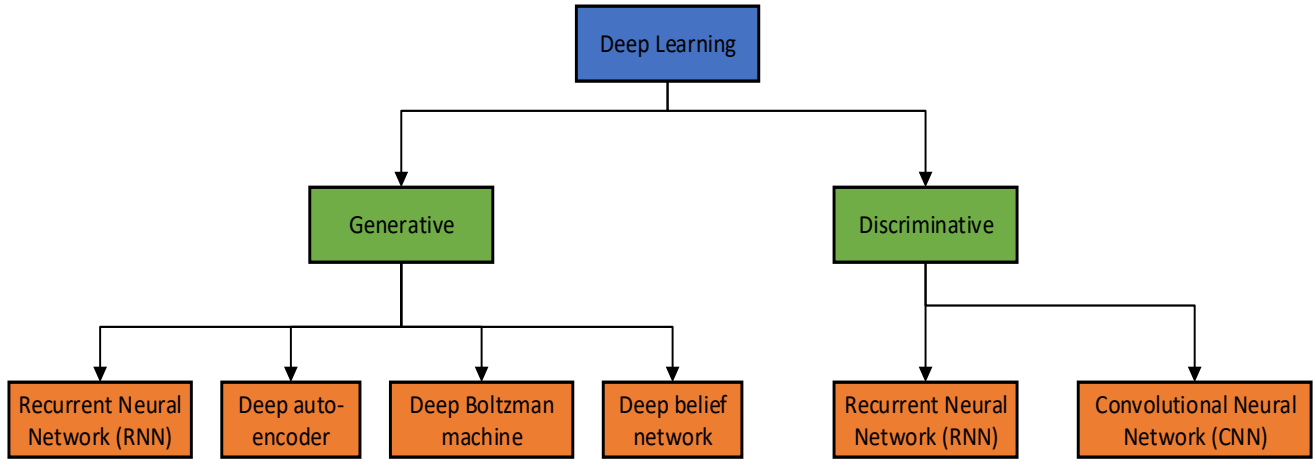

Fig. 4. Classifications of Deep Learning

more attention to the XAI topic since AI has involved many critical sectors. In [13], Miller et al. defined "interpretability" as the degree to which a person understands a decision or prediction of any problem in any field. It is worth noticing that a person might be an expert or a normal person with less or even no prior knowledge in the field.

The term black-box model refers to models whose internal designs are secret or cannot be revealed. Regarding the field of AI, the machine learning model is an example of a black-box model that takes input in the form of image, text or tabular data and produces output without any explanations. XAI aims to illuminate these black-box models by making the models interpretable or explaining their predictions. The term responsible AI is defined in [72] as AI models that consider values, moral and ethical concerns. The authors also propose the ART principles for responsible AI consisting of **Accountability, Responsibility and Transparency**. XAI is the next generation of AI technologies [71] as it shifts AI's development towards designing more robust models, which takes ART principles into account.

1) The need for XAI and its benefits: The ML model is a black-box one; thus, it prevents the development of responsible AI and becomes a barrier to applying AI in further practical implementation. The advancement of such black-box models compromises ART principles and leads to problems such as unethical use, lack of responsibility and accountability, and potential biases in making decisions. Moreover, research in XAI is in utmost need due to laws recently applied by the governments [73]. Arrieta et al. (2020) defined two causes that make the incapability of explaining decisions such a barrier that AI is facing [**?**]. First, in some critical sectors, relying on black-box models to make predictions is impossible due to the enormous gap between the research community and the business operator. Such sectors have such strict regulations, and the decisions to be made daily are so crucial that operators cannot take risks by trusting a model to give vague predictions. Second, the desire to acquire knowledge and improve understanding promotes the development of XAI. The reason is that every field benefited from AI; not only are the results important but also the ability to understand and explain the results.

In [71], XAI's benefits are classified into four different categories, including explain to justify, control, improve and discover. To be more specific, Fig 5 presents the benefits of XAI in a way in which the type of audience is the finest aspect. Firstly, it helps people affected by the model's decisions to be aware of their situation. Thus, they truly understand the decision to comply or even disagree with it. Secondly, XAI benefits experts and users as it provides knowledge and gains trust from them. Thirdly, managers would find it easier to assess regulations with XAI's explanations. Fourthly, regulatory agencies can assess the predictions and explanations to decide whether the model complies with the regulation in force, audits. Lastly, XAI is most useful for developers by helping them ensure and improve the model's efficiency. Data scientists are also benefited from XAI as it provides information for them to create or collect new features to improve performance.

In other words, the appearance of XAI methods can enhance the use of ML techniques for application in different industries by promoting the following benefits:

- **Model Debugging**: During the training process, ML models would take biases from the training dataset, which cannot be preventable. Thus, they tend to provide discriminated decisions against underrepresented groups [74]. XAI can present the patterns constructed by the model, which helps data scientists and developers analyse and erase irrelevant patterns to renovate the dataset.
- **Data Collection**: XAI would give an insight into ML models and a good understanding of the value of the feature in the training set. Thus, the data scientist can evaluate the importance of features and adjust the process of data collecting.
- **Human Decision-making**: In critical sectors where decisions are so important and sensitive, they must be made by humans. XAI can serve as a tool to support humans

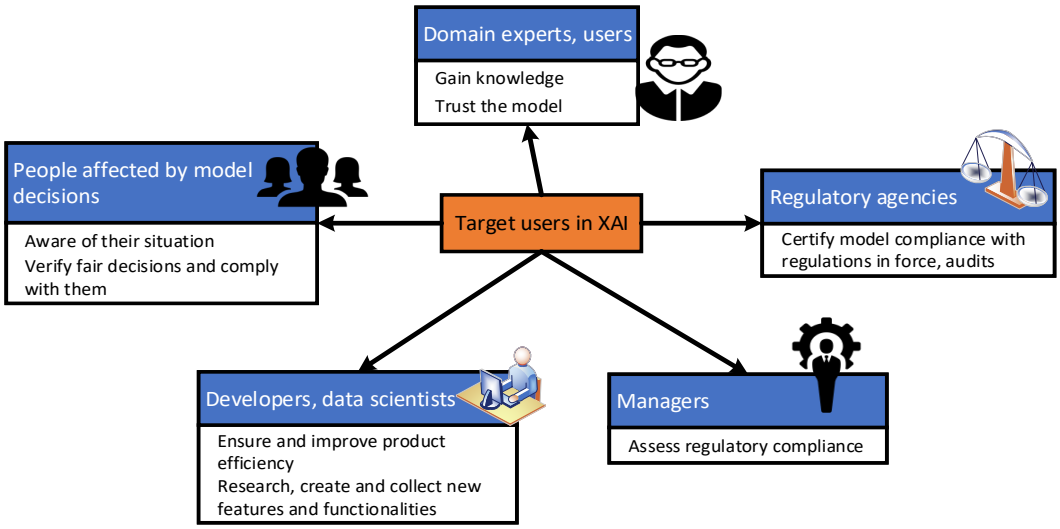

Fig. 5. Different purposes of XAI models for serving various audience profiles

by providing its predictions with reasonable explanations.

- **Trust Building**: This is arguably the ultimate aim of XAI [75] [76]. By getting an insight into the black-box model, people can verify basic facts or identify errors to avoid. Therefore, XAI builds trust between ML models and humans.

In IDS, identifying malicious behaviour is only the first step because understanding such a decision is crucial for a solution. An insight into the decision helps administrators identify the part of the network, the part of features and the security policies compromised by attackers [77]. With the information provided by XAI, the IDS operator can give the correct actions, whether it is to debug the IDS model or apply new security policies to prevent the same attacks in the future. Considering the benefits of XAI discussed above, the need for XAI in IDS is of utmost need.

2) Taxonomy of XAI: Methods for XAI can be classified into different groups based on some criteria:

- Intrinsic or post-hoc: intrinsic methods explain ML models' predictions by restricting the complexity of these models. Meanwhile, post-hoc methods explain the model's results after training by computing their inputs and outputs. The main difference between these two approaches is that most post-hoc approaches can analyze all different ML models. In contrast, the intrinsic approach can only be applied to some specific models whose structures are simple enough to allow intrinsic explainable methods [78].
- Model-specific or model-agnostic: each model-specific method can only be used on a specific model, while model-agnostic methods can be used on many models. When a specific type of explanation is expected, a model-specific approach will limit the choices of black-box models because each model-specific method only provides a type of explanation [71]. Model-specific methods are intrinsic methods, and most post-hoc methods are also model-agnostic methods.

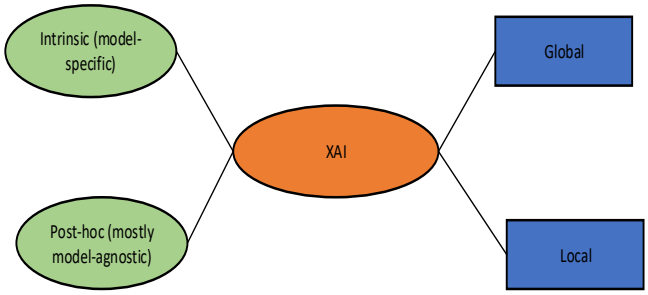

Fig. 6. Classification of XAI methods

- Local explanation or global explanation: while the local explanation methods focus on a single input data instance and utilize different data features to generate explanations, the global explanation methods work on a subset of the whole data instances to summarize the global behaviour of the model [78].

In the scope of this paper, the main focus will be on post-hoc techniques, which are widely used to explain DNN models [**?**]. Nevertheless, first, intrinsically interpretable models will be discussed for comparison.

3) Interpretable models: As discussed above, there are ML models that can be intrinsically interpreted due to their simple structure. This allows researchers and developers to impose constraint directly on the ML model to explain its decision based on internal functioning. A notable example of this approach is proposed in [79]. The authors introduce a generative model named Bayesian Rule Lists based on a decision tree to produce interpretable medical scoring systems. The experiment results show that the proposed model is concise and convincing and suggests applying similar models in other fields.

Xu et al. [80] proposed an attention-based approach for a

model that can learn to describe images. In the experiments using three benchmark datasets, the method has good performance and be able to effectively explain the results to users through visualization [80]. The authors also suggest visual attention for future work. Another work in [81] introduces Supersparse Linear Integer Model (SLIM) for creating data-driven medical scoring systems. Due to the high level of sparsity and small integer coefficients, the model can interpret the results with qualitative understanding [71]. Other interpretable ML models include linear/logistic regression, general additive models, general linear models, decision trees, k-nearest neighbours, and rule-based learners listed in [74], [**?**].

Although this approach provides intuitive explanations that are easy to understand, it can only be applied to models that do not perform very well with a large scale of data. Therefore, the trade-off between explainability and performance hinders this approach's adoption [82]. As discussed previously, deep learning can achieve good performance and flexibility that outperforms traditional machine learning in data with high scale [70]. However, DL's network that consists of multiple layers is considered as a black-box due to the difficulties in explaining its functioning. Therefore, post-hoc approach is utilized in this case to interpret the complex black-box models. This approach can be considered reverse engineering that explains predictions without any modification or knowledge about the model's internal functioning [71].

4) Post-hoc Methods: When it comes to ML models that cannot be interpreted intrinsically due to their sophisticated structures, post-hoc techniques can explain the model after it gives predictions. Most post-hoc methods are model-agnostic which can be applied to any ML model. Such an approach has observed many pieces of research recently due to its desirable advantages as followings [83]:

- Model flexibility: the interpretation techniques are not tied to a specific type of ML model. It is up to developers to choose the most suitable post-hoc methods without changing the black-box model or compromising its high performance.
- Explanation flexibility: the form of explanations is not limited. The post-hoc approach introduces a variety of techniques to generate explanations in different formats. Therefore, the best techniques can be chosen to explain the black-box models depending on the type of audience of XAI's. For example, visualization explanation would be adequate for the lay audience, while feature importance may be more suitable for data scientists.
- Low or no cost to switch: switching the underlying model for a new one is convenient without any modification to the presentation of the explanations. This is a significant advantage over intrinsically interpretable models. Post-hoc methods allow developers to choose the best performing ML model as the underlying model, enhancing accuracy.

Table I lists all the model-agnostic methods that will be discussed. The advantages/disadvantages of each method and their applications are summarized, the table presents useful knowledge about the state-of-the-art techniques.

Visual explanation is the easiest explanation to understand. This technique creates a visualisation of the model's behaviour from its set of inputs and outputs. Visualisations are the most natural way to demonstrate complex interactions within input features or the effect of each feature on the model's prediction to users who might not have expert knowledge about AI techniques such as domain experts and managers. Visualisation-based as a model-agnostic method is a complex task; therefore, it is usually coupled with feature relevance explanations techniques to improve the understanding and provide comprehensive information to the recipient of the XAI model's results [**?**]. A list of representative works using this technique can be found in [105], [106]. In the scope of this paper, some notable techniques of this type will be discussed, including **Partial Dependence Plot** (PDP), **Accumulated Local Effect** (ALE) plot and **Individual Conditional Expectation** (ICE) curves .**Partial Dependence Plot** (PDP) visualize the average marginal effect on the global level of a subset of features on the model's predictions with all other features fixed [107]. Each row of data is considered and predict the outcome by the fitted ML model. Then, the value of the features of interest is altered repeatedly to make a series of predictions.

After iterating through all data instances, the relationship between the output and the features is to visualize. This technique has two major drawbacks: assumption of independence and hidden heterogeneous effect [74]. The term assumption of independence means that there is no relationship between any two features, which is a false assumption. While the hidden heterogeneous effect may occur since the PDP only visualize the average marginal effect. Each data point can have a positive or negative association with the model's prediction; thus, only presenting the average effects of all points would hide such relationships.

A solution to prevent the assumption of independence is the **Accumulated Local Effect** (ALE) plot proposed by Apley et al. (2020). The ALE plot is designed to perform well with highly correlated input features; therefore, there is no need to assume that the features are independent. One more advantage of the ALE plot over PDP is that it requires less computational resource [87]. Goldstein et al. [88] introduce a method called **Individual Conditional Expectation** curves to uncover heterogeneous relationship. ICE plots show one line per data instance, representing the relationship between a feature of interest and the fitted model's decision. Therefore, comparing ICE curves with PDP would produce interesting insights into ML models as ICE curves of each instance might be significantly different from the average of all instances (PDP) [108]. ICE plots also have some limitations. Firstly, the ICE plot is only useful for one feature since drawing two features will create overlaying surfaces [74]. Secondly, ICE curves suffer from the assumption of independence, just like PDP. Lastly, the plot might get overcrowded, hence become impossible to analyze.

Surrogate models are interpretable models that mimic the behaviour of the black-box models. This post-hoc method is especially flexible due to the free choice of surrogate models. However, the generated explanations would not fit all records

TABLE I
COMPARISON OF MODEL-AGNOSTIC METHODS

Model-agnostic methods	Advantages	Disadvantages	References
PDP	- intuitive - causal interpretation	- assumption of independence - hidden heterogenous effects - possibly choose only two features	[84], [85], [86]
ALE plot	- works with correlated features - faster to compute (than PDP)	- complex implementation - are not accompanied by ICE curves	[87]
ICE curve	- intuitive - uncover heterogenous relationships	- possibly display only one feature - assumption of independence - plot can be overcrowded	[88], [89]
Global Surrogate	- flexible - advantages of the model chosen as surrogate	- explanations would not fit all data instances - disadvantages of the model chosen as surrogate	[90], [91], [92]
Local Surrogate (LIME)	- flexible - works for tabular data, text and images	- undefined neighbourhood when applied with tabular data - assumption of independence - explanations can be instable - can be manipulated to hide biases	[75], [93], [94], [95]
Anchors	- easy to understand - works with complex predictions in an instance's neighbourhood	- requires a highly configurable setup - conflicting anchors - complex output spaces - realistic perturbation distributions	[96]
PFI	- highly compressed, global insight - consider all interactions with other features	- assumption of independence - only works with labelled data - results vary greatly when the permutation is repeated.	[97], [98]
SHAP	- effects are distributed fairly - based on a solid theory - allow contrastive explanations	- computationally expensive - assumption of independence.	[99]
Saliency map	- intuitive	- fragile - unreliable	[100]
Counterfactual explanations	- easy to understand, facilitate human reasoning	- multiple counterfactual explanations may contradict each other	[101]
Prototype and criticism	- provide meaningful insights	- may be misleading due to irrelevant features	[102], [103], [104]

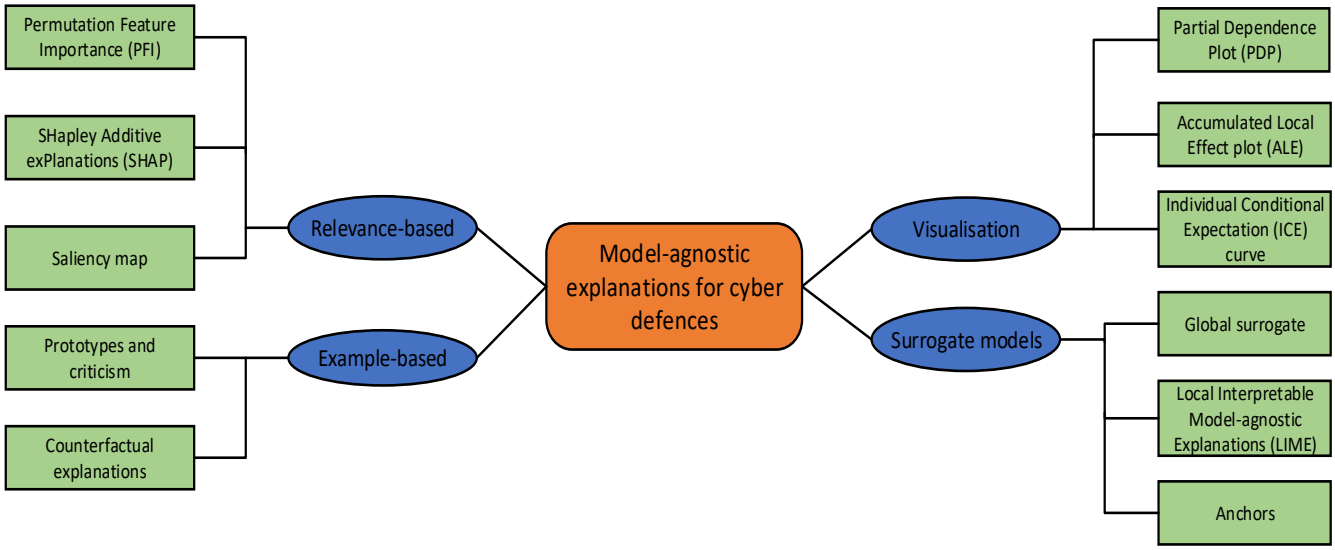

Fig. 7. Classification of XAI-based model-agnostic explanations for cyber defences

in the dataset, and there is no valid theory that supports the surrogate model's representation of the complex model [71]. Moreover, this method suffers from the drawbacks of the interpretable model chosen as the surrogate model. In [75], the author proposed **Local Interpretable Model-agnostic Explanations** (LIME) which derives from the concept of local surrogate models. LIME's key concept is fitting an interpretable model around a specific instance to visualize significant features of that data instance. The interpretable model can be of any type as long as it is a good approximation of the ML model predictions locally. Ribeiro (2016) suggested using LASSO since its performance are the best among linear models [75]. LIME is flexible as it can perform well with tabular data, text and images. However, the correct definition of the neighbourhood is questionable when using LIME with tabular data because improper kernel settings can lead to non-sense explanations [74]. In LIME, data points are sampled from a Gaussian distribution, which means LIME assumes the independence of features. Consequently, invalid data points may be generated and learned by local explanation models.

In [109] the authors show the instability of the explanations produced by LIME. Moreover, Slack et al. (2019) describe

how LIME explanations can be manipulated to hide biases in the dataset [110]. Overall, though LIME is promising, it is in the development phase and needs to be improved a lot before being safely applied. There have been several pieces of research working on fixing LIME's issues or analyze its properties [93], [94], [95]. In [96], the creator of LIME introduces an extension method of LIME using high-precision rules class called **Anchors**. Anchors' explanation is intuitive and easy to understand; however, it requires a highly configurable setup like most other perturbation-based explainers. Ribeiro et al. (2018) also discuss Anchors' limitations, including overly specific anchors, conflicting anchors, complex output spaces, and realistic perturbation distributions.

Relevance-based methods explain the model by ranking the most relevant input features that impact the model's prediction. **Permutation Feature Importance** (PFI) measures the increase in prediction error of a fitted model after permuting the feature's values. This technique shows the importance of each feature by breaking the relationship between the feature and the desired output. Therefore, the model's error would increase if the feature of interest is considered as being important by the model. PFI was first introduced in 2001 by Breiman in [97]. In 2018, Fisher et al. proposed *Model Class Reliance* which is a model-agnostic version of the PFI [98]. PFI provides global insight into the model's behaviour and automatically considers all interactions with other features. PFI requires shuffling the features, which adds randomness to the computation and makes results vary greatly after repeated training. Moreover, PFI suffers from the assumption of independence.

Lundberg et al. [99] propose **SHapley Additive exPlanations** (SHAP) which is based on the game theoretically optimal Shapley Values in [111]. SHAP connects LIME and Shapley values and has several advantages over LIME. The behaviour of the ML model is assumed to be linear locally in LIME, yet the effects are distributed fairly in SHAP. Moreover, SHAP allows contrastive explanations by comparing a prediction to a subset or a single data instance. However, SHAP has to use all the features, hence being computationally expensive. Assumption of independence is also a big problem with SHAP, like many other permutation-based interpretation methods. Another method named saliency maps (or pixel attribution map) [100] is a type of both relevance-based methods and visualization explanations as a pixel of an input image can also be considered as a feature. These approaches generate intuitive explanations as they highlight the most relevant pixels on the final classification. Importance scores of individual pixels are computed using occlusion techniques, or calculations with gradients [71]. However, the saliency map would be fragile [112], and very highly unreliable [113].

Unlike other model-agnostic methods, example-based explanation approaches generate explanations from particular instances of the dataset instead of creating the summaries of the features [71]. Example-based explanation methods have two main techniques, including counterfactual explanations and prototype and criticisms. Wachter et al. [101] introduce counterfactual explanations as a novel model-agnostic XAI method. The technique explains the ML model by describing the change needed in an instance to change or flip the prediction. Explanations generated by this technique is straightforward for human to understand because human usually asks why a specific decision was made instead of other decisions [114]. However, each instance can have multiple counterfactual explanations, and they may contradict each other [74].

Prototypes are representative data instances of the dataset [102], [103] and criticisms are data points that are not well represented by the prototypes [104]. Together with criticism, prototypes can provide meaningful insights into the ML model. Kim et al. [104] develop MMD-critic that selects prototypes and criticism for a dataset to aid human understanding and reasoning. In addition, this technique requires a meaningful data-processing module to select only relevant features because prototypes and criticisms are generated by taking all the features, which may be misleading due to irrelevant features [74].

In the scope of this paper, the two following classes of techniques will be focused on:

- **Feature importance explanations**: the technique is a type of relevance-based method that explains the model's decisions by calculating an importance score for each feature. A comparison among different features' scores would reveal the importance of each feature, which is granted by the model when making decisions [**?**]. A popular XAI post-hoc method using feature relevance techniques mentioned above - SHAP proposed by Lundberg et al. in [99]. The proposed method combines LIME with Shapley value to generate local explanations for the model and demonstrate the relationship between values of input features and the model's decisions.
- **Visual explanations**: this technique utilise the input features and model's prediction to visualise the model's prediction usually by plotting a graph. A visualisation-based method is effective in explaining a black-box model to various recipients who are not expert in AI techniques. Visualisation-based as a model-agnostic method is a complex task; therefore, it is usually coupled with feature relevance explanations techniques to improve the understanding and provide comprehensive information to the recipient of the XAI model's results [**?**].

Based on the concepts discussed above, tools on XAI were developed such as DeepVis Toolbox, TreeInterpreter, Keras-vis, Microsoft InterpretML, MindsDB, SHAP, Tensorboard WhatIf, Tensorflow's Lucid, Tensorflow's Cleverhans and many others. Most of these tools are model-agnostic methods, and a few are model-specific. For instance, DeepVis, kerasvis, and Lucid are for a neural network's explainability, and TreeInterpreter is for a tree-based model's explainability [73]. Each of the proposed approaches has similar concepts at a high level, such as relevance-based, Shapely values, partial dependence plot, surrogate models, counterfactual, prototype, and criticism.

V. Internet of Things and Explainable AI for Cyber Security

In recent years, the term Internet of Things (IoT) is gaining popularity in wireless telecommunications. Its first definition

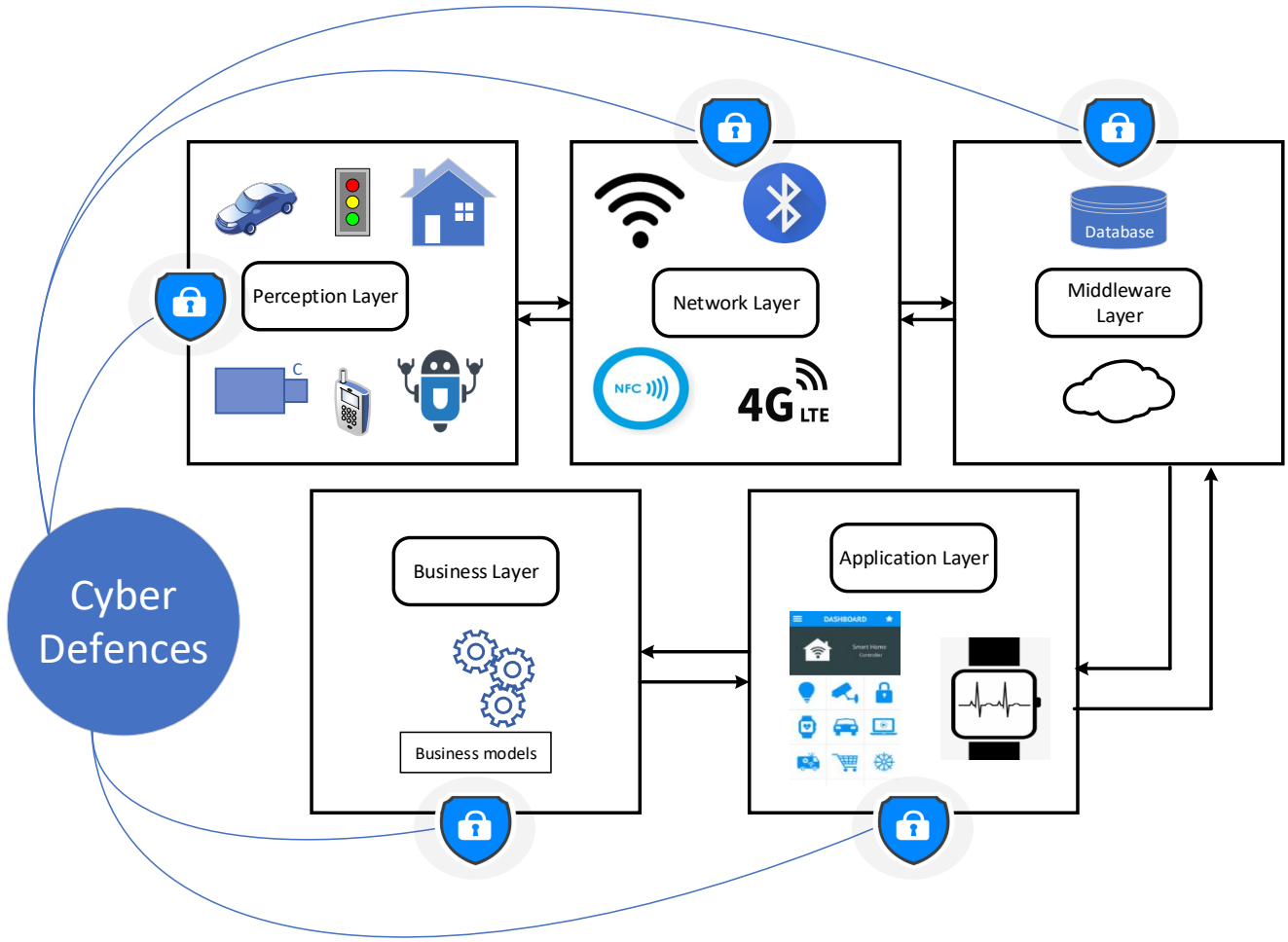

Fig. 8. Systematic architecture of Internet of Things (IoT) for cyber defences

was given by Ashton in the title of a presentation at Procter & Gamble (P&G) in 1999 [115], and the definition directly referred to RFID. Since then, the scope of IoTs has been developed and extended beyond the scope of RFID technologies [116]. According to International Telecommunication Union (ITU), IoT is a global infrastructure for the information society which enables services by interconnecting things based on existing and evolving information and communication technologies [117]. At the same time, alternative definitions have been proposed to emphasize different subjects such as connected things in IoT, Internet-related aspects of IoT, semantic challenges in the IoT and many others.

Generally, the term IoT refers to the network of physical devices, objects, vehicles, buildings embedded with sensors, software, and other communication technologies to collect, store, analyse and exchange data with other devices and systems. Evolving technologies, which build IoT, are characterized by the development of low-powered, low-cost processors, wireless networking, artificial intelligence, and mobile computing [118]. Together, these technologies make communication between people, processes, and things become much more accessible, improving the efficiency and quality of human life. Therefore, technologies related to IoT has been advanced rapidly, making it the fastest-growing technology in computing [3]. By the year 2025, IoT and related applications are estimated to have an economic impact of $3.9 trillion to $11.1 trillion per year [119].

The domain of IoT typically consists of a wide range of advanced technologies; thus, there is no single reference architecture that can represent all possible implementation of IoT networks. However, for research purposes, the five-layer IoT architecture is utilized to focus on the finest aspects of IoT which can fulfill the requirements of security and privacy, as well as assist to develop and evaluate new applications of cyber defences, such as intrusion detection and threat hunting [120]. Figure 8 shows the five-layer architecture of IoT. The perception layer is the sensor layer that collects information through the sensors attached to physical objects such as cars, robots, surveillance cameras, phones and many others. The network layer connects devices and servers and transmits collected data from the perception layer using Wifi, Bluetooth, Near-Field-Communication (NFC) and other methods. Middleware layer process and analyze transportation data. It purifies the data and only extracts useful information. Cloud computing and big

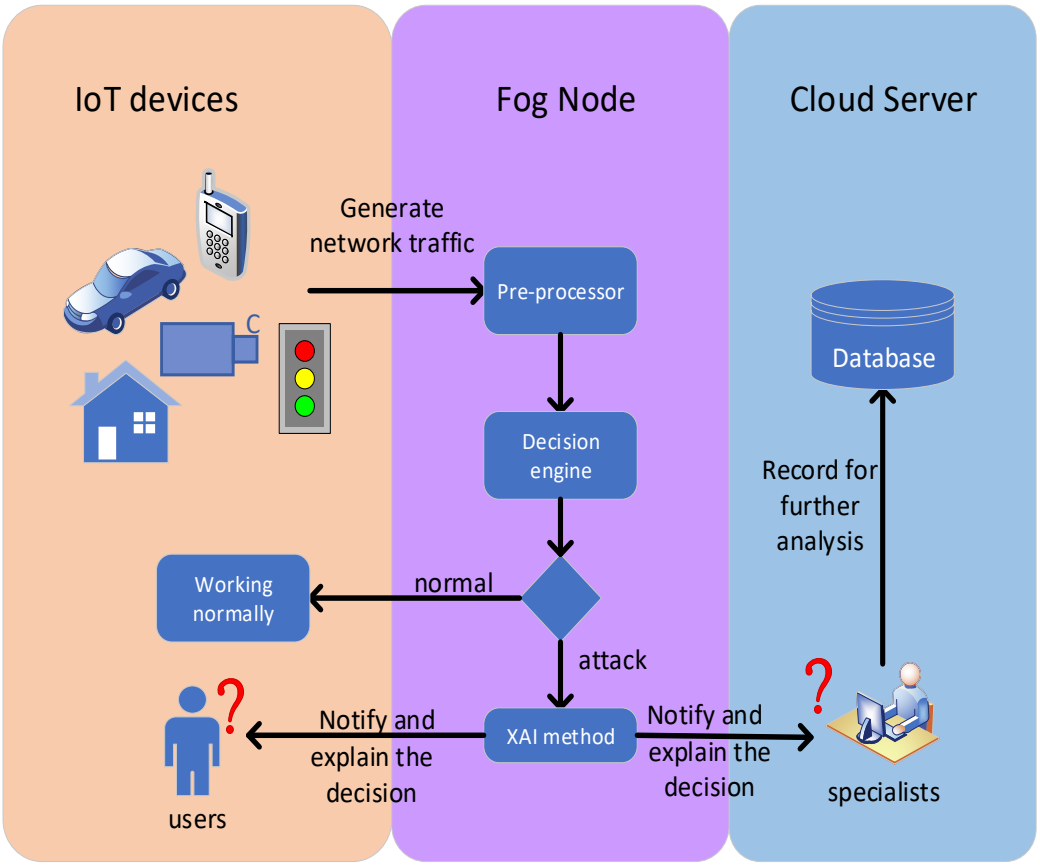

Fig. 9. Architecture of explainable AI-based Cyber defences in IoT networks

data processing modules are examples of technologies used in this layer [121]. Application layer refers to IoT applications that provide services to users such as smart home, smart health and many others. Lastly, the business layer is responsible for controlling IoT applications, business models and user privacy [122].

IoT application tasks are always delegated to cloud computing due to the limited computational resources on IoT devices. Cloud receives and handle IoT data to process, then sends back signals to IoT devices for further actions. This approach is also referred as Cloud of Things (CoT) paradigm and it is flexible and robust in managing IoT devices. However, this centralised architecture normally suffers from high communication latency and power consumption; thus, large-scale deployment of CoT is almost impossible in real life [123]. With the emergence of fog computing (also known as edge computing), the conventional approach was replaced. Fog computing is an extension of cloud computing that distributes the benefits of the cloud closer to the IoT and across five layers of the IoT [1]. Fog nodes can be any devices with computing, storage and network connectivity such as industrial controllers, switches routers and surveillance cameras. Generally, the fog was developed to address two major problems in IoT network including high network latency and sensitive data leakage. IoT applications are written for fog nodes at the network edge, allowing fog nodes to ingest IoT data from multiple devices. Then, different types of IoT data is directed to a proper place for analysis based on the level of time-sensitive; the places can be fog nodes, fog aggregation nodes or cloud [124]. Therefore, fog nodes offer an advantage for IoT network by allowing distributed security services due to their ability to offload computational tasks from IoT devices [125].

In recent years, several pieces of research [45], [126] has focused on utilising fog computing to design a distributed IDS in IoT system. IDS would be trained in a cloud environment then executed on a fog node; thus it prevents the latency problem apparent in centralised-based solutions. Moreover, fog nodes can analyse and explain each network activity generated by IoT devices as they are directly connected to sensor devices [1]. An AI-based IDS that is deployed at fog node would be very useful for processing and obtaining hidden information from big IoT data [127]. In regards to the advantages of this approach, we recommend utilising fog computing to design a distributed explainable AI-based IDS. Additionally, different explanations generated from the IDS will be directed to the optimal place for analysis and inform multiple audiences. Figure 9 illustrates the architecture of model for anomaly-based IDS in IoT network. First of all, telemetry data from various IoT devices are sent to fog nodes for pre-processing. Then, an ML-based decision engine classifies the activity as

normal or attack. Normal activities are allowed to happen. In contrast, abnormal traffic is terminated and sent to the XAI model for processing. The model generates different types of explanation that is directed to inform audiences or being stored in a cloud server for further research.

A. Cyber threats in IoT networks

IoT has appeared in all aspects of society, including industry, healthcare, homes, sport, peer-to-peer networks [128], entertainment, and others [129]. It is gradually replacing many conventional computer systems in those fields. However, in comparison with conventional computing systems, this engagement has introduced new security challenges for several reasons: [3]:

- **Complex and diverse environments**: an IoT is diverse as it is connected with a wide variety of devices, platforms, communication means and protocols. The diversity enhances the usability and convenience of IoT technologies yet at the cost of numerous potential targets and attack vectors.
- **Undefined boundary**: The IoT system does not have a well-defined boundary since it changes very often due to the mobility of users and devices. Due to the undefined boundary, it is not easy to design effective security mechanisms for the IoT system. Moreover, the IoT systems would suffer from an extensive attack surface.
- **Connection between virtual and physical systems**: In unattended working environments, IoT devices would be a potential target since the attacker can access them physically. Moreover, IoT devices can function on the received data, which optimises the connection between virtual and physical systems yet allows the attacker to convert the potential physical consequences quickly.
- **Limited energy and computational resources on devices**: Most IoT devices have limited energy and computational resources, making it hard to implement decentralised IDS and advanced security techniques on physical objects. For example, IDS that utilises deep neural networks require GPU to perform well in real-time, which most IoT devices cannot afford.

Regarding the reasons discussed above, IoT networks are attractive targets for cyber attacks. Many attack techniques are utilised to exploit the security issues at different layers of the IoT environment, including sensing, network, middleware, application and business layer. According to [130], 41% of attacks exploit IoT devices' vulnerabilities; then, victim devices can be utilized to launch a large-scale attack. For example, a Mirai-based attack compromised a French cloud computing company in 2016 and became the most significant distributed denial of service attack (DDoS) recorded at that time [4]. Attackers were reported to instrument zombified IoT devices and use them as a pivot point to launch a DDoS attack on the French web host. Researchers then blame default, and weak security configurations [3]. This example has testified to the secure authentication mechanisms and traffic classification techniques. Thus, to prevent cyber threats in IoT systems, new defence mechanisms should be developed. IDS is considered as the primary method to attain these requirements [131].

TABLE II
PERFORMANCE COMPARISON [132]

Algorithm	Accuracy	Precision	Recall	f1-score
DNN (1 layers)	0.929	0.998	0.915	0.954
DNN (2 layers)	0.929	0.998	0.914	0.954
DNN (3 layers)	0.930	0.997	0.915	0.955
DNN (4 layers)	0.929	0.999	0.913	0.954
DNN (5 layers)	0.927	0.998	0.911	0.953
Ada Boost	0.925	0.995	0.911	0.951
Decision Tree	0.928	0.999	0.912	0.953
K-Nearest Neighbour	0.929	0.998	0.913	0.954
Linear Regression	0.848	0.989	0.821	0.897
Naive Bayes	0.929	0.988	0.923	0.955
Random Forest	0.927	0.999	0.910	0.953
SVM* (linear)	0.811	0.994	0.770	0.868
SVM* (rbf)	0.811	0.992	0.772	0.868

VI. ML-BASED CYBER DEFENCES

In recent years, the use of ML algorithms and techniques in holistic cyber defences, such as IDS, threat intelligence, threat hunting, privacy preservation, digital forensics, has become common. Researchers produced several research studies related to the applications of ML and cyber defences, with a focus on IDS in this study. Rahul et al. [132] used the KDDCup-99 dataset to compare the performance of several classical ML algorithms and Deep Neural Networks (DNN) models, which have from 1 to 5 layers. After running 100 epochs, the DNN model with three layers outperformed all other DNN models and classical ML algorithms. Results, shown in Table II, demonstrate the efficiency of deep learning models in IDSs.

Different from other detection methods, DL-based IDS must consider handling overfitting and model optimization. Overfitting means that the model works well on training data but it has poor performance on unknown records; therefore, it is ineffective in real life. Model optimization aims to minimize a loss function to improve the effectiveness of the model by applying optimisers such as Stochastic gradient descent (SGD), Adam and others. In regards to those characteristics and a few others including accuracy rate and performance comparison, Table III summarises the advantages and disadvantages of existing literature in DL-based IDSs.

KDDCup-99 was used in [133] to train CNN (convolutional neural network), CNN-RNN, CNN-LSTM and CNN-GRU. The experiment was run up to 1000 epochs with a learning rate in the range [0.01-0.05], and results were recorded. The complex network structure suffered from overfitting and was outperformed by simple network structures. Results showed that CNN 1 layer obtains the best accuracy (0.999), precision (0.999), recall (0.999) and F-score (0.999). This paper applied CNN for IDS by modelling the network traffic events as time-series of TCP/IP packets. From the experiment results, the authors claimed that modelling network activities as series of TCP/IP packets are an efficient method to train DL models such as CNN, RNN, LSTM, or GRU.In [134] and [135], the authors also used KDDCup-99 to train LSTM-RNN for

TABLE III
DEEP LEARNING-BASED IDS

Reference	Techniques	Dataset	Advantages	Disadvantages
[132]	DNN(1-5 layers), classic ML	KDD-CUP99	- good comparison between multiple techniques	- no model optimization - outdated and unbalanced dataset
[133]	CNN with CNN-LSTM with CNN-GRU	KDD-CUP99	- experiment with different configurations	- overfitting - no model optimization - outdated and unbalanced dataset
[134]	LSTM-RNN	KDD-CUP99	- high detection rate - experiment with different configurations	- outdated and unbalanced dataset
[135]	LSTM-RNN	KDD-CUP99	- high detection rate - experiment with different configurations	- overfitting - outdated and unbalanced dataset
[136]	RNN with Hessian-free optimization	KDD-CUP99	- low false alarm rate	- overfitting - outdated and unbalanced dataset
[137]	GRU with MLP	KDD-CUP99 NLS-KDD	- comparison between two datasets - good accuracy	- do not claim which specific dataset the model has been tested on - overfitting - no optimization - outdated dataset
[138]	RNN	NLS-KDD	- comparison between RNN and classic ML	- overfitting - no optimization - outdated dataset
[139]	GRU-RNN with CNN with RF	NLS-KDD	- good accuracy - short training time - low computational resources - experiment on different configurations	- outdated dataset
[140]	RNN, ANN, classic ML	NLS-KDD	- comparison between multiple techniques	- do not provide testing results on challenging dataset KDDTest-21 - outdated dataset - overfitting
[141]	LSTM with feature embedding, MLP, ML	UNSW-NB15	- comparison between multiple techniques - feature embedding	- no optimization
[142]	Bi-LSTM	UNSW-NB15	- feature selection (only consider 5 features)	- only use a small subset of UNSW-NB15 - overfitting - no optimization
[143]	DNN	UNSW-NB15, CIDDS, GPRS	- comparison between different datasets	- no optimization - lack comparison between multiple techniques
[144]	LSTM, ML	CIDDS	- good comparison between multiple techniques	- overfitting

anomaly-based IDS, and both papers obtained the best detection rate of 98.95%.

Xu et al.[137] combined multiple techniques to build an IDS using KDDCUP-99 and NSL-KDD datasets. In the proposed model, GRUs are utilized to build the main memory unit, and MLP was used to classify intrusions. The results showed the detection rate of 99.42% and 99.31% on KDDCup-99 and NSL-KDD, respectively. However, the authors did not claim which specific dataset the model was tested on, neither KDDTrain+, KDDTest+, nor KDDTest-21. Yin et al. (2017) proposed an RNN model and conducted testing by changing the number of hidden neurons and learning rate to find the optimal hyperparameters [138]. The RNN model and other models using conventional ML algorithms were tested on KDDTest+ and KDDTest-21. Results showed that RNN outperformed other methods in both binary and multiclass classification. In the binary classification, RNN achieved an accuracy of 83.28% and 68.55% in KDDTest+ and KDDTest-21, respectively. In the multi-class classification, the accuracy was 81.29% and 64.67% in KDDTest+ and KDDTest-21, respectively.

Andalib and Vakili [139] proposed an IDS combining three different types of ML techniques (GRU-RNN, CNN and RF) using the NSL-KDD dataset for training and testing. Results showed that the proposed model can achieve good accuracy on KDDTest+ and KDDTest-21, which were 87.28% and 76.61%, respectively. Moreover, Andalib and Vakili (2020) claimed that the ML model has a short training time and only needs low computational resources. Chaibi et al. (2020) proposed an architecture implemented as two methodologies, then trained and tested the model on the NSL-KDD dataset for five class attack categories [140]. Results showed that RNN outperformed ANN and ANN outperformed other ML classifiers. However, the paper has the identical drawback with the paper in [137] where the authors did not provide performance on the challenging testing dataset - KDDTest-21.

Gwon et al. (2019) proposed an LSTM with feature embedding to build an IDS and utilized UNSW-NB15 for training and testing. The experiment assumed that the data instances follow a timely order, which means that the dataset captures temporal dependence for intrusion detection [141]. LSTM models with feature embedding can obtain the accuracy of 99.72% for binary classification and 86.98% for multiclass classification, which outperformed MLP and other conventional ML algorithms. The authors contributed an excellent performance to embedding techniques. Due to the distinguishable information of features in the UNSW-NB15 dataset, feature embedding is an efficient way to capture such information to design an effective model.

Roy and Cheung [142] built a Bi-Directional LSTM-RNN for anomaly-based IDS and obtained an accuracy rate of 95.71% with a false alarm rate of 0%. However, the model was only trained and tested on a small subset of the UNSW-NB15

dataset. Tama and Rhee [143] proposed an IDS built upon DNN model for IoT network using new datasets including UNSW-NB15, CIDDS-001 and GPRS. The model performed perfectly on CIDDS-001 in which it achieves about 100% accuracy, precision and recall. However, the paper lacks a performance comparison between DNN and other ML algorithms. The research in [144] used the CIDDS dataset to train and test the LSTM model. The performance was compared between different ML algorithms, and LSTM outperformed all other methods with an accuracy of 84.83%.

A. DL-based IDS for IoT Networks

The Bot-IoT dataset [145] is an IoT benchmark dataset suffering from an imbalanced data problem in which it has a small amount of benign data and a large amount of attack data. In regards to the solutions for balancing this dataset and other characteristics, table IV summarises the advantages and disadvantages of recent literature in DL-based IDS for IoT.

Ferreg and Maglaras [146] proposed a DeepCoin framework built upon blockchain and deep learning. The proposed model utilized RNN for IDS and evaluated the performance on three datasets including the Bot-IoT dataset and two others. The authors balanced the dataset before using it for training the model. Although model optimization was not used, the IDS achieves a good accuracy of 98.20% on the Bot-IoT dataset. Aldhaheri et al. [147] develop Deep Learning and Dendritic Cell Algorithm (DeepDCA) to design an IDS. The proposed model extracts features from Bot-IoT then uses it for training and evaluating. The authors use a balanced version of the dataset but do not provide information about extracting the dataset.

The creators of the Bot-IoT dataset also use the created dataset to test RNN and LSTM [145]. Before training the model, they pre-process the dataset by calculating the correlation coefficient among the features and use normalization to scale features' value within the range [0, 1]. To solve the imbalanced data problem in the Bot-IoT dataset, authors in [148] and [149] use Synthetic Minority Oversampling Technique (SMOTE) before training the proposed IDS built upon ANN and Temporal CNN, respectively. The TCNN model shows promising results with low training time; however, it was not implemented in online mode for real-time detection.

Ge et al. [150] propose a feed-forward neural networks model for binary and multi-class classification of intrusions in IoT devices. They use a balanced version of the Bot-IoT dataset and Adam as the model optimizer. The performance is good; however, the model only works in batch mode. Moreover, the authors use a list of well-known ports for encoding port columns. Booij et al. [151] evaluated multiple classifiers on two IoT datasets including ToN_IoT and Aposemat IoT-23. They conducted a novel approach of cross-training by using data fusion on the data level and observed that training on one dataset and testing with the other provide bad results. Therefore, the inclusion of configurations is crucial. In the experiment, they used default parameter settings for training neural network; therefore, the neural network is outperformed by Gradient Boosting Machine (GBM) and Random Forest (RF).

Moustafa [152] presented a novel testbed architecture that was used to collect heterogeneous data sources from IoT/IIOT devices, Windows and Linux-based operating systems, as well as network traffic. 4 different classifiers were built and evaluated by using a small part of the ToN_IoT network dataset. Moustafa used the IP addresses and ports when building models to collect benchmark outputs for comparison purposes. It is recommended to exclude IP addresses and ports in the data features.

B. Explainable AI-based IDS

To the best of our knowledge, there are only a few work-related to XAI in IDS. Most of the XAI's works focus on fields such as computer vision and natural language processing. An attempt of designing an XAI model in IDS is proposed in [77]. The authors apply a decision tree to build an ML model, then train and evaluate the performance on the KDD benchmark dataset. Because the decision tree is an interpretable model, it can explain the decisions using feature engineering, and the rule-based model. Although the decision tree model outperforms logistic regression and support vector machine in predicting the classes between malicious and normal behaviours, its performance is not compared with the deep learning model.

Islam et al. (2019) proposed a method to gather and utilize domain knowledge to automate the defence response and improve the explainability of the IDS model. In the experiment, domain knowledge (i.e., CIA principles) is instilled into the ML models, and the CIDS2017 dataset is utilized [153]. Results from the experiment show an increase in generalizability and explainability of the models, which builds trust in the IDS model and opens the door to adapt to big data from numerous IoT devices [153].

Marino et al. (2018) produced explanations for incorrect samples by using an adversarial approach [154]. It finds the minimum modification to correctly label the incorrect records then visualize the most critical features to explain the model's wrong decision. The authors experiment on the NSL-KDD99 dataset. Some advantages of the approach include (1) being a model-agnostic method, (2) require no modification to the internal structure of the black-box model and (3) being extendable for further analysis of the model [154].

Le et al. [155] propose a IDS built upon an RNN and explain the outcomes of the model to generate Software Defined Networking (SDN) flow rules. They use a linear regression model as a local surrogate describes in LEMNA [156], the k most relevant features are chosen to generate network access control policies. The authors used NSL-KDD as a benchmark dataset.

VII. Research Challenges and Future Direction

A. Challenges of building an efficient NADS in IoT networks

Despite many pieces of research in NADS showing good performance with benchmarking datasets, it is challenging to build an effective IDS with a high detection rate, scalability, robustness and protection against all attack vectors [3], especially zero-day attacks and multi-stage attacks. This

TABLE IV
DEEP LEARNING-BASED IDS FOR IOT

Reference	Techniques	Dataset	Advantage	Disadvantage
[146]	RNN	Bot-IoT	- balanced dataset - high accuracy	- no model optimization
[147]	DeepDCA	Bot-IoT	- high accuracy - loss function as model optimizer	- no information about how dataset is balanced
[145]	RNN, LSTM	Bot-IoT	- compute correlation coefficient among features	- overfitting - unbalanced dataset - no model optimization
[148]	ANN	Bot_IoT	SMOTE to balance dataset	- overfitting - no model optimization
[149]	Temporal CNN	Bot-IoT	SMOTE to balance dataset	- online mode was not implemented for real time detection
[150]	feed-forward neural network	Bot-IoT	balanced dataset	- model only works in batch mode - poor encoding technique for port columns
[151]	GBM, RF, NN	ToN_IoT Aposemat IoT-23	- novel approach of cross-training - comparison between IoT datasets - high accuracy	- default parameter settings for training NN model
[152]	GBM, RF, NB, DNN	ToN_IoT network	- novel testbed architecture - high accuracy	- only a small part of the dataset was used - used IP addresses and ports to build models

section presents challenges for designing an efficient IDS in IoT networks.

- Data source is a key component of NADS in both the training and testing process. However, it is impossible to construct a dataset that involves all normal and malicious behaviours in a heterogeneous environment of IoT networks. Many existing datasets suffer from missing labels and poor attack diversity. Moreover, most of them collect an incomplete set of features, and network information is captured without including both headers and payloads [14]. In addition, the IDS may only perform well on a limited number of devices because collected data only contains network activities of a few types of IoT devices [3].
- Real-time detection is also a challenge. Collecting and monitoring network traffic in real-time would cause a long processing time and a high false alarm rate, which can degrade the network performance significantly. Therefore, the data processing module and the detection method must be adopted carefully to mitigate this [14].
- Although the IDS using DL-based techniques shows good performance, apply such techniques in IoT environments is difficult due to the limited computation resources on IoT devices. Moreover, ML and DL-based techniques are computationally expensive, which leads to network latency issues and hence, become impossible to be used in critical sectors such as the internet of vehicles (IoVs) [157].
- Defending against multi-stage attacks remains a challenge for existing IDS techniques. IDS effectively detects single-stage attacks as malicious activities are conducted over a short period. The MSA utilizes more complex attack techniques over a long period [31] in which each step taken is insufficient to be recognized as aggression. Therefore, the MSA avoids being detected by IDS. In [158], the authors point out several challenges for detecting MSA, including: (1) modelling MSAs, (2) building a system to detect and track the progress of interleaved MSA, and (3) constructing datasets with interleaved MSA scenarios.

B. Challenges of developing effective XAI-based Cyber Defences

While the area of XAI is being focused on and efforts have been put on, many limitations have to be improved to obtain explainability in AI. This section presents challenges for XAI-based cyber defences.

- **Accessible Explanations**: instead of focusing on the demands of different types of audience as in Fig 5, most current XAI methods produce results that only make sense to researchers. In addition, some XAI methods produce explanations as feature important vectors, which is inherently low-level explanations. This format would be useful for developers and researchers to debug or design the models. However, other audience types may find such explanation format complex, confusing and useless [15]. Ideally, the XAI method should provide accessible explanations for all types of audience, which means they should be easily utilised by society, especially policymakers and the law [**?**]. This remains a big challenge in the field of XAI.
- **Standardised Terminology and Evaluation of Explanations**: the XAI research community does not have a standardised terminology. Terminologies in XAI is defined differently by researchers as there are no standardised definitions and vocabularies [**?**], which would lead

to confusion. This makes it impossible to evaluate and compare the performance of different XAI methods as the ground truth is unknown. Moreover, no criteria have been standardised that consider the subjective measures used in human-centred evaluations. It is also impossible to evaluate all XAI methods by a defined evaluation metric [159]. To develop a practical evaluation of XAI explanations, all the above aspects need to be considered carefully.

- **Statistical Uncertainty**: Many XAI methods (PFI, SHAP) are subject to uncertainty as they provide explanations by computing from data. However, the uncertainty of the explanation is not addressed as explanations are given. Consequently, the explanations are not reliable and compromise the responsible AI. To fix this, a rigorous approach should be adopted to study the uncertainty of XAI methods [160]. Otherwise, XAI has to face statistical testing problems such as p-hacking [161].
- **Systematic Instability**: This characteristic affects the performance of XAI methods that analyse the internal structure of the models. Many different ML models can perform well on a specifically labelled dataset; however, their internal pathway might differ due to the complexity of ML techniques. Such differences would lead to changes in generated explanations across multiple models [71].
- **Feature Dependence**: Many XAI methods (PDP, LIME, PFI) suffer from the assumption of independence; thus, they automatically create invalid data points. Using such data points probably degrades the reliability of XAI methods as explanations are given unreal data instances and will never happen in real-life [160]. Technically, when features in the dataset are correlated, the explanations would be misleading.

To sum up, the field of explainable artificial intelligence (XAI) emerged from the need to understand the predictions and functioning of machine learning (ML) models, and to develop robust XAI models that benefit various types of audience [**?**]. XAI methods have witnessed many studies and applications across a wide range of subjects, such as image recognition [162], [163], [164] and natural language processing [165]. However, there have been very few studies in intrusion detection systems (IDS). Nowadays, as human life significantly depends on computer systems, IDS is crucial for the security of any organisation's system. Additionally, there is a growing need for explainable AI-based IDS since the model's recipients require explanations for many valid purposes. Throughout the literature, there are many examples of ML-based IDS (subsection **??**, VI-A) and various XAI methods (subsection IV-C) with very few addressing explainable AI-based IDS (subsection VI-B).

This paper identifies a gap in existing research that is of utmost importance. The field of XAI is relatively young, and there is limited existing literature addressing the explainable AI-based IDS. There is a lack of studies in creating new XAI methods for IDS and the evaluations of existing XAI approaches for ML-based IDS. This provides a research opportunity, to develop an understanding of existing XAI methods for IDS and design a novel explainable AI-based IDS. Current challenges in this research field were addressed in section VII and we highly recommend future research to tackle those existing problems as well as consider those challenges when designing a new XAI method or anomaly-based IDS.

VIII. Conclusion

The paper has examined the importance, current challenges, and recent works in network security, Intrusion Detection Systems (IDS), Artificial Intelligence (AI) and security issues in the Internet of Things (IoT) networks. Since digital transformation is happening across many industries, the security and privacy of computer systems arise as a big concern. Intrusion Detection System is a prominent method to protect cyberspace due to its convenience and automation without the need for human action. Different types of IDS can fit a wide range of organisations depending on the IDS placement and detection method. Recent literature shows that anomaly-based IDS, which uses machine learning algorithms and deep learning as an underlying detection method, achieves excellent results in preventing unknown attacks. Moreover, the heterogeneous nature of IoT also encourages the utilization of AI techniques in IDS due to AI's capability of analysing and learning attack patterns from a large scale of data. This paper has surveyed recent works in ML-based IDS, especially DL-based IDS. It is observed that many proposed models suffer from outdated or unbalanced datasets, which can degrade the model's efficiency in real-life implementation. In the future, this work will assist researchers to explore the challenges and future studies of explainable AI for developing context-aware anomaly detection methods in IoT networks

References

[1] K. Tange, M. De Donno, X. Fafoutis, and N. Dragoni, "A systematic survey of industrial internet of things security: Requirements and fog computing opportunities," *IEEE Communications Surveys Tutorials*, vol. 22, no. 4, pp. 2489–2520, 2020.

[2] J. Srinivas, A. K. Das, and N. Kumar, "Government regulations in cyber security: Framework, standards and recommendations," *Future Generation Computer Systems*, vol. 92, pp. 178–188, 2019.

[3] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, "From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3369–3388, 2018.

[4] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for internet of things (iot) security," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020.

[5] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: the confluence of edge computing and artificial intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020.

[6] J. Navarro, A. Deruyver, and P. Parrend, "A systematic survey on multi-step attack detection," *Computers & Security*, vol. 76, pp. 214–249, 2018.

[7] I. Ghafir, M. Hammoudeh, V. Prenosil, L. Han, R. Hegarty, K. Rabie, and F. J. Aparicio-Navarro, "Detection of advanced persistent threat using machine-learning correlation analysis," *Future Generation Computer Systems*, vol. 89, pp. 349–359, 2018.

[8] Y.-D. Lin, "Editorial: Second quarter 2020 ieee communications surveys and tutorials," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 790–795, 2020.

[9] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," *Journal of information security and applications*, vol. 44, pp. 80–88, 2019.

[10] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions," *SN Computer Science*, vol. 2, no. 3, pp. 1–18, 2021.

[11] Y.-D. Lin, "Editorial: Third quarter 2020 ieee communications surveys and tutorials," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1466–1471, 2020.

[12] J. D. Moore and W. R. Swartout, "Explanation in expert systemss: A survey," UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, Tech. Rep., 1988.

[13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[14] N. Moustafa, J. Hu, and J. Slay, "A holistic review of Network Anomaly Detection Systems: A comprehensive survey," *Journal of Network and Computer Applications*, vol. 128, pp. 33–55, 2 2019.

[15] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, p. 68–77, Dec. 2019. [Online]. Available: https://doi.org/10.1145/3359786

[16] S. P. Portillo, J. Manuel Estévez, and T. Leganés, "Attacks against intrusion detection networks: evasion, reverse engineering and optimal countermeasures," Universidad Carlos III de Madrid, Tech. Rep., 2014. [Online]. Available: https://core.ac.uk/download/pdf/29406661.pdf

[17] S. Morgan, "Cybercrime to cost the world $10.5 trillion annually by 2025," *Cybercrime Magazine*, vol. 13, 2020.

[18] ACSC, *Australian Cyber Security Centre Annual Cyber Threat Report 2020-21*, 2021. [Online]. Available: https://www.cyber.gov.au/acsc/view-all-content/reports-and-statistics/acsc-annual-cyber-threat-report-2020-21

[19] S. Das, "Direct costs associated with cybersecurity incidents costs australian businesses $29 billion per annum," 2018. [Online]. Available: https://news.microsoft.com/en-au/features/direct-costs-associated-with-cybersecurity-incidents-costs-australian-businesses-29-billion-per-annum/

[20] E. G, "Us government to spend over $18 billion on cybersecurity - atlas vpn," 2020.

[21] B. von Solms and R. von Solms, "Cybersecurity and information security – what goes where?" *Information and Computer Security*, vol. 26, no. 1, pp. 2–9, 2018.

[22] D. Kuehl, "From cyberspace to cyberpower: Defining the problem." *Cyberpower and National Security*, vol. 30, no. National Defense University Press, Washington, D.C, 2009.

[23] D. J. Gunkel, "Hacking cyberspace," *JAC*, pp. 797–823, 2000.

[24] H. Ghadeer, "Cybersecurity issues in internet of things and countermeasures," in *2018 IEEE International Conference on Industrial Internet (ICII)*, 2018, pp. 195–201.

[25] M. Abomhara and G. M. Kien, "Cyber Security and the Internet of Things: Vulnerabilities, Threats, Intruders and Attacks," *Journal of Cyber Security and Mobility*, vol. 4, no. 1, pp. 65–88, 1 2015.

[26] S. Mittal, P. K. Das, V. Mulwad, A. Joshi, and T. Finin, "CyberTwitter: Using Twitter to generate alerts for cybersecurity threats and vulnerabilities," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*. Institute of Electrical and Electronics Engineers Inc., 11 2016, pp. 860–867.

[27] H. Zhang, P. Cheng, L. Shi, and J. Chen, "Optimal denial-of-service attack scheduling with energy constraint," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 3023–3028, 11 2015.

[28] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.

[29] T. Rid and B. Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies*, vol. 38, pp. 4–37, 1 2015. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01402390.2014.977382

[30] J. Zhao, S. Shetty, J. W. Pan, C. Kamhoua, and K. Kwiat, "Transfer learning for detecting unknown network attacks," *EURASIP Journal on Information Security*, vol. 2019, no. 1, pp. 1–13, 2019.

[31] J. Shin, S.-H. Choi, P. Liu, and Y.-H. Choi, "Unsupervised multi-stage attack detection framework without details on single-stage attacks," *Future Generation Computer Systems*, vol. 100, pp. 811–825, 2019.

[32] L. Martin, "Cyber kill chain." [Online]. Available: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[33] I. Ghafir, V. Prenosil *et al.*, "Advanced persistent threat attack detection: an overview," *Int J Adv Comput Netw Secur*, vol. 4, no. 4, p. 5054, 2014.

[34] R. Bace and P. Mell, "Intrusion Detection Systems," National Institute of Standards and Technology (NIST), Tech. Rep., 2001.

[35] R. Wazirali, "An improved intrusion detection system based on knn hyperparameter tuning and cross-validation," *Arabian Journal for Science and Engineering*, vol. 45, no. 12, pp. 10 859–10 873, 2020.

[36] P. Sharma, J. Sengupta, and P. Suri, "Survey of intrusion detection techniques and architectures in cloud computing," *International Journal of High Performance Computing and Networking*, vol. 13, no. 2, pp. 184–198, 2019.

[37] S. A. Ludwig, "Intrusion detection of multiple attack classes using a deep neural net ensemble," in *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, vol. 2018-Janua, 2 2018, pp. 1–7.

[38] H. A. Kholidy and F. Baiardi, "Cids: A framework for intrusion detection in cloud systems," in *2012 Ninth International Conference on Information Technology - New Generations*, 2012, pp. 379–385.

[39] L. Santos, C. Rabadao, and R. Gonçalves, "Intrusion detection systems in internet of things: A literature review," in *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, 2018, pp. 1–7.

[40] D. Moon, S. B. Pan, and I. Kim, "Host-based intrusion detection system for secure human-centric computing," *The Journal of Supercomputing*, vol. 72, no. 7, pp. 2520–2536, 2016.

[41] M. D. Singh, "Computer Network and Information Security," *Computer Network and Information Security*, vol. 8, pp. 41–47, 2014. [Online]. Available: http://www.ossec.net/files/ossec-hids-

[42] J. Peng, K.-K. R. Choo, and H. Ashman, "User profiling in intrusion detection: A review," *Journal of Network and Computer Applications*, vol. 72, pp. 14–27, 2016.

[43] M. Kumar, M. Hanumanthappa, and T. V. Suresh Kumar, "Encrypted traffic and ipsec challenges for intrusion detection system," in *Proceedings of International Conference on Advances in Computing*, A. Kumar M., S. R., and T. V. S. Kumar, Eds. New Delhi: Springer India, 2012, pp. 721–727.

[44] E. Besharati, M. Naderan, and E. Namjoo, "Lr-hids: logistic regression host-based intrusion detection system for cloud environments," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 9, pp. 3669–3692, 2019.

[45] M. A. Rahman, A. T. Asyhari, L. Leong, G. Satrya, M. H. Tao, and M. Zolkipli, "Scalable machine learning-based intrusion detection system for iot-enabled smart cities," *Sustainable Cities and Society*, vol. 61, p. 102324, 2020.

[46] M. F. Elrawy, A. I. Awad, and H. F. Hamed, "Intrusion detection systems for iot-based smart environments: a survey," *Journal of Cloud Computing*, vol. 7, no. 1, pp. 1–20, 2018.

[47] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, 2017.

[48] J. Ribeiro, F. B. Saghezchi, G. Mantas, J. Rodriguez, and R. A. Abd-Alhameed, "Hidroid: Prototyping a behavioral host-based intrusion detection and prevention system for android," *IEEE Access*, vol. 8, pp. 23 154–23 168, 2020.

[49] J. Ribeiro, F. B. Saghezchi, G. Mantas, J. Rodriguez, S. J. Shepherd, and R. A. Abd-Alhameed, "An autonomous host-based intrusion detection system for android mobile devices," *Mobile Networks and Applications*, vol. 25, no. 1, pp. 164–172, 2020.

[50] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *Elsevier*, 2012.

[51] N. Moustafa, G. Misra, and J. Slay, "Generalized outlier gaussian mixture technique based on automated association features for simulating and detecting web application attacks," *IEEE Transactions on Sustainable Computing*, pp. 1–1, 2018.

[52] T. Saranya, S. Sridevi, C. Deisy, T. D. Chung, and M. K. Khan, "Performance Analysis of Machine Learning Algorithms in Intrusion Detection System: A Review," in *Procedia Computer Science*, vol. 171. Elsevier B.V., 2020, pp. 1251–1260.

[53] S. Dua and X. Du, *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[54] W. Haider, G. Creech, Y. Xie, and J. Hu, "Windows based data sets for evaluation of robustness of host based intrusion detection systems (ids) to zero-day and stealth attacks," *Future Internet*, vol. 8, no. 3, p. 29, 2016.

[55] A. Vasudevan, E. Harshini, and S. Selvakumar, "Ssenet-2011: A network intrusion detection system dataset and its comparison with

kdd cup 99 dataset," in *2011 Second Asian Himalayas International Conference on Internet (AH-ICI)*, 2011, pp. 1–5.

[56] N. Moustafa, G. Creech, and J. Slay, *Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models*. Cham: Springer International Publishing, 2017, pp. 127–156.

[57] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," vol. 46, no. 4, 2014.

[58] D. Li, L. Deng, M. Lee, and H. Wang, "Iot data feature extraction and intrusion detection system for smart cities based on deep migration learning," *International journal of information management*, vol. 49, pp. 533–545, 2019.

[59] H. S. Anderson, A. Kharkar, B. Filar, and P. Roth, "Evading machine learning malware detection," *Black Hat*, 2017.

[60] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, "Deep learning for unsupervised insider threat detection in structured cybersecurity data streams," *arXiv preprint arXiv:1710.00811*, 2017.

[61] Q. Chen, Q. Liao, Z. L. Jiang, J. Fang, S. Yiu, G. Xi, R. Li, Z. Yi, X. Wang, L. C. Hui *et al.*, "File fragment classification using grayscale image conversion and deep learning in digital forensics," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 140–147.

[62] X. Wang, J. Zhang, A. Zhang, and J. Ren, "Tkrd: Trusted kernel rootkit detection for cybersecurity of vms based on machine learning and memory forensic analysis," *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2650–2667, 2019.

[63] X.-D. Zhang, "Machine learning," in *A Matrix Algebra Approach to Artificial Intelligence*. Springer, 2020, pp. 223–440.

[64] Y. Liu, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Blockchain and machine learning for communications and networking systems," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 1392–1431, 2020.

[65] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 2018, pp. 322–327.

[66] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[67] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[68] S. Raschka and V. Mirjalili, *Python Machine Learning - Second Edition*. Packt Publishing, 2017.

[69] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Shallow and deep networks intrusion detection system: A taxonomy and survey," *arXiv*, 1 2017.

[70] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv*, 1 2019.

[71] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[72] V. Dignum, "Responsible artificial intelligence: designing ai for human values," 2017.

[73] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable Artificial Intelligence Approaches: A Survey," *arXiv*, 1 2021.

[74] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[75] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[76] B. Kim, E. Glassman, B. Johnson, and J. Shah, "ibcm: Interactive bayesian case model empowering humans via intuitive interaction," 2015.

[77] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, 2021.

[78] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *ArXiv*, vol. abs/2006.11371, 2020.

[79] B. Letham, C. Rudin, T. H. McCormick, D. Madigan *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[80] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[81] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349–391, 2016.

[82] S. Sarkar, T. Weyde, A. Garcez, G. G. Slabaugh, S. Dragicevic, and C. Percy, "Accuracy and interpretability trade-offs in machine learning applied to safer gambling," in *CEUR Workshop Proceedings*, vol. 1773. CEUR Workshop Proceedings, 2016.

[83] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[84] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[85] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *Journal of Business & Economic Statistics*, vol. 39, no. 1, pp. 272–281, 2021.

[86] B. M. Greenwell, "pdp: An r package for constructing partial dependence plots." *R J.*, vol. 9, no. 1, p. 421, 2017.

[87] D. Apley, "Visualizing the effects of predictor variables in black box supervised learning models. arxiv," *arXiv preprint arXiv:1612.08468*, 2016.

[88] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[89] A. Goldstein, A. Kapelner, and J. Bleich, "Icebox: Individual conditional expectation plot toolbox," 2017.

[90] N. Puri, P. Gupta, P. Agarwal, S. Verma, and B. Krishnamurthy, "Magix: Model agnostic globally interpretable explanations," *arXiv preprint arXiv:1706.07160*, 2017.

[91] Y. Ming, H. Qu, and E. Bertini, "Rulematrix: Visualizing and understanding classifiers with rules," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 342–352, 2018.

[92] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," *arXiv preprint arXiv:1711.09784*, 2017.

[93] L. Hu, J. Chen, V. N. Nair, and A. Sudjianto, "Locally interpretable models and effects based on supervised partitioning (lime-sup)," *arXiv preprint arXiv:1806.00663*, 2018.

[94] J. Rabold, H. Deininger, M. Siebers, and U. Schmid, "Enriching visual with verbal explanations for relational concepts–combining lime with aleph," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 180–192.

[95] A. H. A. Rahnama and H. Boström, "A study of data and label shift in the lime framework," *arXiv preprint arXiv:1910.14421*, 2019.

[96] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[97] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[98] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019.

[99] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[100] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[101] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[102] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *The Annals of Applied Statistics*, pp. 2403–2424, 2011.

[103] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 260–269.

[104] B. Kim, R. Khanna, and O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 2288–2296.

[105] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences*, vol. 225, pp. 1–17, 2013.

[106] P. Cortez and M. J. Embrechts, "Opening black box data mining models using sensitivity analysis," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2011, pp. 341–348.

[107] A. Stojić, N. Stanić, G. Vuković, S. Stanišić, M. Perišić, A. Šoštarić, and L. Lazić, "Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition," *Science of The Total Environment*, vol. 653, pp. 140–147, 2019.

[108] M. Kłosok, M. Chlebus *et al.*, *Towards Better Understanding of Complex Machine Learning Models Using Explainable Artificial Intelligence (XAI): Case of Credit Scoring Modelling*. University of Warsaw, Faculty of Economic Sciences, 2020.

[109] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[110] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "How can we fool lime and shap? adversarial attacks on post hoc explanation methods," 2019.

[111] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[112] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[113] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, *The (Un)reliability of Saliency Methods*. Cham: Springer International Publishing, 2019, pp. 267–280.

[114] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.

[115] K. Ashton *et al.*, "That 'internet of things' thing," *RFID journal*, vol. 22, no. 7, pp. 97–114, 2009.

[116] F. Wortmann and K. Flüchter, "Internet of things," *Business & Information Systems Engineering*, vol. 57, no. 3, pp. 221–224, 2015.

[117] I. T. Union, "Internet of things global standards initiative," 2012.

[118] F. Firouzi and B. Farahani, *Architecting IoT Cloud*. Cham: Springer International Publishing, 2020, pp. 173–241.

[119] J. Diechmann, K. Heineke, T. Reinbacher, and D. Wee, "The internet of things: How to capture the value of iot," Tech. Rep., 2018.

[120] M. Burhan, R. A. Rehman, B. Khan, and B.-S. Kim, "Iot elements, layered architectures and security issues: A comprehensive survey," *Sensors*, vol. 18, no. 9, 2018.

[121] M. A. J. Jamali, B. Bahrami, A. Heidari, P. Allahverdizadeh, and F. Norouzi, "Iot architecture," *Towards the Internet of Things*, pp. 9–31, 2020.

[122] M. Mukherjee, I. Adhikary, S. Mondal, A. K. Mondal, M. Pundir, and V. Chowdary, "A vision of iot: applications, challenges, and opportunities with dehradun perspective," in *Proceeding of International Conference on Intelligent Communication, Control and Devices*. Springer, 2017, pp. 553–559.

[123] Y.-D. Lin, "Editorial: Fourth quarter 2020 ieee communications surveys and tutorials," *IEEE Communications Surveys Tutorials*, vol. 22, no. 4, pp. 2130–2135, 2020.

[124] Cisco, "Fog computing and the internet of things: extend the cloud to where the things are," Tech. Rep., 2016.

[125] S. Venticinque and A. Amato, "A methodology for deployment of iot application in fog," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 5, pp. 1955–1976, 2019.

[126] P. Kumar, G. P. Gupta, and R. Tripathi, "A distributed ensemble design based intrusion detection system using fog computing to protect the internet of things networks," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2020.

[127] D. C. Nguyen, P. Cheng, M. Ding, D. Lopez-Perez, P. N. Pathirana, J. Li, A. Seneviratne, Y. Li, and H. V. Poor, "Enabling ai in future wireless networks: A data life cycle perspective," *IEEE Communications Surveys Tutorials*, vol. 23, no. 1, pp. 553–595, 2021.

[128] R. Pecori, "A pki-free key agreement protocol for p2p voip applications," in *2012 IEEE International Conference on Communications (ICC)*, 2012, pp. 6748–6752.

[129] H. Tahaei, F. Afifi, A. Asemi, F. Zaki, and N. B. Anuar, "The rise of traffic classification in iot networks: A survey," *Journal of Network and Computer Applications*, vol. 154, p. 102538, 2020.

[130] U. 42, "2020 unit 42 iot threat report," Palo Alto, Tech. Rep., 2020. [Online]. Available: https://unit42.paloaltonetworks.com/iot-threat-report-2020/

[131] T. Marsden, N. Moustafa, E. Sitnikova, and G. Creech, "Probability risk identification based intrusion detection system for scada systems," 2017.

[132] V. K. Rahul, R. Vinayakumar, K. Soman, and P. Poornachandran, "Evaluating Shallow and Deep Neural Networks for Network Intrusion Detection Systems in Cyber Security," in *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*. Institute of Electrical and Electronics Engineers Inc., 10 2018.

[133] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1222–1228.

[134] J. Kim, J. Kim, H. L. Thi Thu, and H. Kim, "Long short term memory recurrent neural network classifier for intrusion detection," in *2016 International Conference on Platform Technology and Service (PlatCon)*, 2016, pp. 1–5.

[135] T. Le, J. Kim, and H. Kim, "An effective intrusion detection classifier using long short-term memory with gradient descent optimization," in *2017 International Conference on Platform Technology and Service (PlatCon)*, 2017, pp. 1–6.

[136] J. Kim and H. Kim, "Applying recurrent neural network to intrusion detection with hessian free optimization," in *Information Security Applications*, H.-w. Kim and D. Choi, Eds. Cham: Springer International Publishing, 2016, pp. 357–369.

[137] C. Xu, J. Shen, X. Du, and F. Zhang, "An intrusion detection system using a deep neural network with gated recurrent units," *IEEE Access*, vol. 6, pp. 48 697–48 707, 2018.

[138] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21 954–21 961, 2017.

[139] A. Andalib and V. T. Vakili, "An autonomous intrusion detection system using an ensemble of advanced learners," in *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, 2020, pp. 1–5.

[140] N. Chaibi, B. Atmani, and M. Mokaddem, "Deep learning approaches to intrusion detection: A new performance of ann and rnn on nsl-kdd," in *Proceedings of the 1st International Conference on Intelligent Systems and Pattern Recognition*, ser. ISPR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 45–49.

[141] H. Gwon, C. Lee, R. Keum, and H. Choi, "Network intrusion detection based on lstm and feature embedding," *ArXiv*, vol. abs/1911.11552, 2019.

[142] B. Roy and H. Cheung, "A deep learning approach for intrusion detection in internet of things using bi-directional long short-term memory recurrent neural network," in *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*, 2018, pp. 1–6.

[143] B. Adhi Tama and K. H. Rhee, "Attack classification analysis of iot network via deep learning approach," *Research Briefs on Information Communication Technology Evolution (ReBICTE)*, vol. 3, 11 2017.

[144] S. A. Althubiti, E. M. Jones, and K. Roy, "Lstm for anomaly-based network intrusion detection," in *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*, 2018, pp. 1–3.

[145] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[146] M. A. Ferrag and L. Maglaras, "Deepcoin: A novel deep learning and blockchain-based energy exchange framework for smart grids," *IEEE Transactions on Engineering Management*, vol. 67, no. 4, pp. 1285–1297, 2020.

[147] S. Aldhaheri, D. Alghazzawi, L. Cheng, B. Alzahrani, and A. Al-Barakati, "Deepdca: novel network-based detection of iot attacks using artificial immune system," *Applied Sciences*, vol. 10, no. 6, p. 1909, 2020.

[148] Y. N. Soe, P. I. Santosa, and R. Hartanto, "Ddos attack detection based on simple ann with smote for iot environment," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, pp. 1–5.

[149] A. Derhab, A. Aldweesh, A. Z. Emam, and F. A. Khan, "Intrusion detection system for internet of things based on temporal convolution neural network and efficient feature engineering," *Wireless Communications and Mobile Computing*, vol. 2020, 2020.

[150] M. Ge, X. Fu, N. Syed, Z. Baig, G. Teo, and A. Robles-Kelly, "Deep learning-based intrusion detection for iot networks," in *2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC)*, 2019, pp. 256–25 609.

[151] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. H. den Hartog, "Ton_iot: The role of heterogeneity and the need for

standardization of features and attack types in iot network intrusion datasets," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[152] N. Moustafa, "A new distributed architecture for evaluating ai-based security systems at the edge: Network ton_iot datasets," *Sustainable Cities and Society*, vol. 72, p. 102994, 2021.

[153] S. R. Islam, W. Eberle, S. K. Ghafoor, A. Siraj, and M. Rogers, "Domain Knowledge Aided Explainable Artificial Intelligence for Intrusion Detection and Response," *arXiv*, 11 2019.

[154] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 3237–3243.

[155] H. Li, F. Wei, and H. Hu, "Enabling dynamic network access control with anomaly-based ids and sdn," in *Proceedings of the ACM International Workshop on Security in Software Defined Networks Network Function Virtualization*, ser. SDN-NFVSec '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 13–16.

[156] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "Lemna: Explaining deep learning based security applications," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 364–379.

[157] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, "Iot security techniques based on machine learning: How do iot devices use ai to enhance security?" *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 41–49, 2018.

[158] T. Shawly, M. Khayat, A. Elghariani, and A. Ghafoor, "Evaluation of hmm-based network intrusion detection system for multiple multi-stage attacks," *IEEE Network*, vol. 34, no. 3, pp. 240–248, 2020.

[159] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, 2021.

[160] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges," *arXiv preprint arXiv:2010.09337*, 2020.

[161] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science," *PLOS Biology*, vol. 13, no. 3, pp. 1–15, 03 2015.

[162] I. Kim, S. Rajaraman, and S. Antani, "Visual interpretation of convolutional neural network predictions in classifying medical image modalities," *Diagnostics*, vol. 9, no. 2, p. 38, 2019.

[163] S. Shi, X. Zhang, and W. Fan, "Explaining the predictions of any image classifier via decision trees," *arXiv preprint arXiv:1911.01058*, 2019.

[164] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," *arXiv preprint arXiv:1807.08024*, 2018.

[165] S. M. Mathews, "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review," in *Advances in Intelligent Systems and Computing*, vol. 998. Springer Verlag, 7 2019, pp. 1269–1292.

Nour Moustafa is Coordinator of Postgraduate Cyber Discipline & Leader of Intelligent Security at School of Engineering & Information Technology (SEIT), University of New South Wales (UNSW)'s UNSW Canberra, Australia. He was a Post-doctoral Fellow at UNSW Canberra from June 2017 till December 2018. He received his Ph.D. degree in the field of Cyber Security from UNSW Canberra in 2017. He obtained his Bachelor and Master degree of Computer Science in 2009 and 2014, respectively, from the Faculty of Computer and Information, Helwan University, Egypt. His areas of interest include Cyber Security, in particular, Network Security, IoT security, intrusion detection systems, statistics, Deep learning and machine learning techniques. He has several research grants with totalling over AUD 1.2 Million. He has been awarded the 2020 prestigious Australian Spitfire Memorial Defence Fellowship award. He is also a Senior IEEE Member, ACM Distinguished Speaker, as well as CSCRC and Spitfire Fellow. He has served his academic community, as the guest associate editor of IEEE transactions journals, including IEEE Transactions on Industrial Informatics, IEEE IoT Journal, as well as the journals of IEEE Access, Future Internet and Information Security Journal: A Global Perspective. He has also served over seven conferences in leadership roles, involving vice-chair, session chair, Technical Program Committee (TPC) member and proceedings chair, including 2020–2021 IEEE TrustCom and 2020 33rd Australasian Joint Conference on Artificial Intelligence.

Nam Pham is currently an Honours student at School of Engineering & Information Technology (SEIT), University of New South Wales (UNSW)'s UNSW Canberra, Australia. He obtained his Bachelor of Computing and Cyber Security from UNSW Canberra in 2020. His areas of interest include Cyber Security, IoT security, Intrusion Detection System, Deep Learning, Machine Learning techniques and Explainable Artificial Intelligence.

Izhar Ahmed Khan received the PhD degree in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), the master's degree in computer science from Mid Sweden University,Sweden, and the B.Sc. degree from the University of Engineering and Technology, Pakistan. He is currently a postdoctoral fellow at computer science with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include machine learning, intrusion detection, anomaly detection systems and Explainable Artificial Intelligence.

Albert Y. Zomaya is Chair Professor of High-Performance Computing Networking in the School of Computer Science and Director of the Centre for Distributed and High-Performance Computing at the University of Sydney. To date, he has published more than 600 scientific papers and articles and is (co-)author/editor of more than 30 books. A sought-after speaker, he has delivered 250 keynote addresses, invited seminars, and media briefings. His research interests span several areas in parallel and distributed computing and complex systems. He is currently the Editor in Chief of the ACM Computing Surveys and served in the past as Editor in Chief of the IEEE Transactions on Computers (2010-2014) and the IEEE Transactions on Sustainable Computing (2016-2020).

Professor Zomaya is a decorated scholar with numerous accolades including Fellowship of the IEEE, the American Association for the Advancement of Science, and the Institution of Engineering and Technology (UK). Also, he is an Elected Fellow of the Royal Society of New South Wales and an Elected Foreign Member of Academia Europaea. He is the recipient of the1997 Edgeworth David Medal from the Royal Society of New South Wales for outstanding contributions to Australian Science, the IEEE Technical Committee on Parallel Processing Outstanding Service Award (2011), IEEE Technical Committee on Scalable Computing Medal for Excellence in Scalable Computing (2011), IEEE Computer Society Technical Achievement Award (2014), ACM MSWIM Reginald A. Fessenden Award (2017), and the New South Wales Premier's Prize of Excellence in Engineering and Information and Communications Technology (2019).