KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com
ScienceDirect

ICT Express 10 (2024) 935–958

www.elsevier.com/locate/icte

Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects

Iqbal H. Sarker[a,b,*], Helge Janicke[a,b], Ahmad Mohsin[a,b], Asif Gill[c], Leandros Maglaras[d]

[a] *Centre for Securing Digital Futures, Edith Cowan University, Perth, WA, 6027, Australia*
[b] *Cyber Security Cooperative Research Centre, Perth, WA, 6027, Australia*
[c] *University of Technology Sydney, Sydney, Australia*
[d] *Edinburgh Napier University, Edinburgh, UK*

Received 17 February 2024; received in revised form 25 April 2024; accepted 18 May 2024
Available online 21 May 2024

Abstract

Digital twins (DTs) are an emerging digitalization technology with a huge impact on today's innovations in both industry and research. DTs can significantly enhance our society and quality of life through the virtualization of a real-world physical system, providing greater insights about their operations and assets, as well as enhancing their resilience through real-time monitoring and proactive maintenance. DTs also pose significant security risks, as intellectual property is encoded and more accessible, as well as their continued synchronization to their physical counterparts. The rapid proliferation and dynamism of cyber threats in today's digital environments motivate the development of automated and intelligent cyber solutions. Today's industrial transformation relies heavily on artificial intelligence (AI), including machine learning (ML) and data-driven technologies that allow machines to perform tasks such as self-monitoring, investigation, diagnosis, future prediction, and decision-making intelligently. However, to effectively employ AI-based models in the context of cybersecurity, human-understandable explanations, and their trustworthiness, are significant factors when making decisions in real-world scenarios. This article provides an extensive study of explainable AI (XAI) based cybersecurity modeling through a taxonomy of AI and XAI methods that can assist security analysts and professionals in comprehending system functions, identifying potential threats and anomalies, and ultimately addressing them in DT environments in an intelligent manner. We discuss how these methods can play a key role in solving contemporary cybersecurity issues in various real-world applications. We conclude this paper by identifying crucial challenges and avenues for further research, as well as directions on how professionals and researchers might approach and model future-generation cybersecurity in this emerging field.

Keywords: Cybersecurity; Explainable AI; Machine learning; Data-driven; Automation; Intelligent decision-making; Trustworthiness; Digital twin

1. Introduction

Digital twins (DTs) are a virtual representation of a physical entity or system that uses data and Artificial Intelligence (AI) to simulate and analyze its behavior, performance, and other characteristics [1,2]. DT can be used in a wide range of industries, including manufacturing [3], smart cities [4], critical services [5], healthcare [6], agriculture [7], energy [8], and so on to improve efficiency, optimize processes, and reduce costs. DT can provide organizations with useful insights extracted from data about their operations and assets. However, DT can also be vulnerable to cyber threats such as unauthorized access, data breaches, or other malicious attacks as DTs become increasingly interconnected with other systems and devices through the Internet of Things (IoT) and other smart technologies, discussed briefly in Section 3. Cybersecurity threats can impact the confidentiality, integrity, and availability of DT data, as well as the safety and reliability of the physical system being represented by DT. Therefore, it is crucial to take into account an automated and intelligent cybersecurity systems design that meets today's needs.

Recent advancements in AI, including machine learning (ML) methods, significantly changed how we might combine and analyze data, and eventually apply the extracted insights or knowledge for automation and intelligent decision-making processes in various real-world application areas [9]. By gathering massive amounts of data and effectively analyzing it to

* Corresponding author at: Centre for Securing Digital Futures, Edith Cowan University, Perth, WA, 6027, Australia.
E-mail address: m.sarker@ecu.edu.au (I.H. Sarker).
Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2024.05.007

identify harmful patterns and unusual behaviors, AI technologies have become essential to the cybersecurity industry [10]. However, to use AI-based models effectively in the context of cybersecurity in DT, human-understandable explanations, and their trustworthiness, are considered significant factors when making decisions in real-world scenarios. Thus the key aspects are:

- *Automation:* It involves the use of automated processes, algorithms, and tools to streamline and enhance cybersecurity tasks within a digital twin environment. With the automation of repetitive and time-consuming security tasks, organizations can detect and respond to security threats more rapidly, reduce manual work and human error, and free up staff resources for more strategic security initiatives.
- *Intelligence:* Typically, intelligence refers to the capability of learning, understanding, and applying knowledge for intelligent decision-making to perform tasks that require human intelligence. Thus, it includes analyzing and interpreting enormous amounts of data produced within the digital twin ecosystem. The discovered knowledge from data enables organizations to gain insight into emerging threats, detect anomalous behavior, identify potential vulnerabilities, predict cyber threats, and generate actionable insights to improve security.
- *Trustworthiness:* This encompasses the reliability, integrity, and credibility of the security mechanisms and processes implemented within the digital twin environment. Transparency in implementing and validating cybersecurity measures, as well as accountability for security incidents and breaches, contribute to trustworthiness. It is crucial to establish trustworthiness in cybersecurity within digital twin ecosystems to promote confidence among stakeholders and ensure their resilience and sustainability.

Overall, the key aspects for cybersecurity modeling in a digital twin ecosystem are automation, *i.e.*, reducing manual efforts with self-learning, intelligence, *i.e.*, informed decision-making based extracted insights, and trustworthiness, *i.e.*, human-interpretable cyber decisions, which enable efficient and effective protection against evolving threats in increasingly complex digital environments. Thus, a trade-off among "Automation", "Intelligence", and "Trustworthiness", representing "CyberAIT" in short, is important as shown in Fig. 1. A more transparent and understandable AI model, also known as XAI, could therefore make cybersecurity modeling more effective. In a DT environment, analysts and security professionals can use this information to comprehend how the system operates, identify potential vulnerabilities and threats, and ultimately make the best actionable decisions to successfully address them. A motivational scenario highlighting the significance of XAI has also been presented in Fig. 2. Taking into account the key aspects of "CyberAIT", this paper focuses on AI and XAI-based methods for cybersecurity modeling with their potential real-world applications.

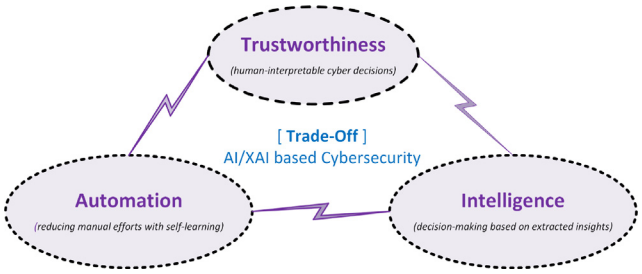

Fig. 1. An illustration of the key aspects — Automation (A), Intelligence (I) and Trustworthiness (T) of today's Cybersecurity (CyberAIT).

1.1. Related surveys and our contributions

Throughout the last few years, surveys on XAI have been typically conducted with an emphasis on Black-Box models (internal workings and decision-making processes are opaque and difficult to interpret by humans), *e.g.*, deep neural network-based modeling. For instance, Adadi et al. [12] presented a survey on XAI peeking inside the Black-Box. Similarly, Ibrahim et al. [13], and Guidotti et al. [14] presented XAI focusing Black-Box systems, methods, and relevant applications in their survey. Recently, Dwivedi et al. [15] explored XAI in terms of approaches, programming frameworks, and software toolkits in more detail. For the academic and industrial communities, these surveys offer fundamental knowledge and valuable lessons. There is still a need for a succinct exposition of AI's use in cybersecurity and digital twin, though. There have been some literary efforts regarding XAI for cybersecurity, although they have tended to concentrate on particular goals. For example, Capuano et al. [16] and Alcaraz et al. [2] presented XAI focusing on various cyber threats and approaches in the context. More related works are summarized in Table 2. However, an extensive study on AI/XAI-based modeling with their explainable capabilities by taking into account "CyberAIT" needs to be explored to comprehend its potential real-world use cases in the context of cybersecurity in the digital twin. Thus we have formulated five key questions below to understand the main focus of this paper, which are needed to answer and discuss to make this paper beneficial for the cyber and DT community:

(i) Is modern cybersecurity modeling in DT required to be automated and intelligent with trustworthy decisions?
(ii) Does AI-based cybersecurity modeling including machine learning and data-driven technologies have the potential to meet today's diverse security concerns in DT?
(iii) What aspects and characteristics do AI and XAI-based methodologies have that make the decision-making process human-understandable and resolve today's cyber issues in DT more effectively?
(iv) What are the diverse real-world usage potentials within the context of cybersecurity in digital twin and how AI/XAI-based methods can lead to?
(v) What are the key challenges of AI-based cyber modeling, and how can scientists and researchers overcome the issues in this emerging area of study?

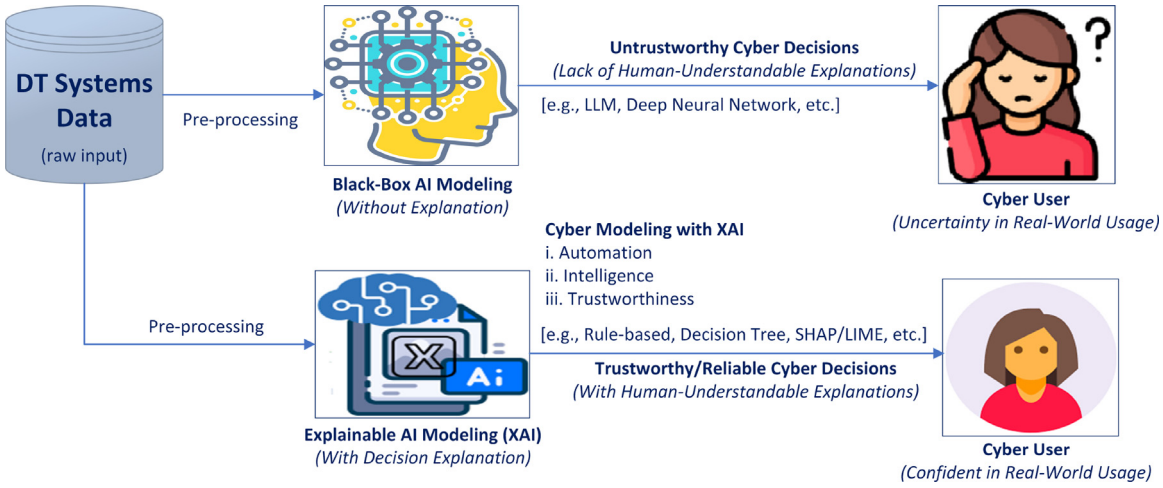

Fig. 2. A motivational scenario highlighting Black-Box AI modeling (internal workings and decision-making processes are opaque and difficult to interpret by humans) vs XAI-based cyber modeling from the perspective of a cyber user (adopted from Sarker et al. [11]).

In this paper, we aim to answer these important questions from the perspective of AI and XAI-based cybersecurity modeling with their potentiality in various real-world use cases, which builds the foundation of our contribution. To the best of our knowledge, this study is the first effort to provide a thorough synthesize, analysis, and discussion about the AI/XAI-based cybersecurity modeling exploring multi-aspects intelligence such as machine learning, deep learning, data-driven, rule discovery, semantic knowledge, as well as multimodal intelligence for future enhancements of cyberspace by taking into account CyberAIT.

Overall, our specific contributions are as follows:

- We survey and compare the existing literature to identify the scope of our paper regarding cybersecurity automation, intelligence, and trustworthiness in the DT environment.
- We emphasize DT for enhancing cyber resilience. We also highlight possible threats and anomalies in DT that are needed to mitigate. For this, we explore diverse functional layers from the physical to the application layer of a digital twin with associated cyber issues and the necessity to employ AI/XAI-based cybersecurity modeling.
- We present a taxonomy of AI/XAI-based cybersecurity modeling methods and discuss their computing capabilities and potential. Our discussion also focuses on making them human-understandable in a cybersecurity context.
- We explore how AI/XAI-based cybersecurity models can be used in real-world applications, ranging from anomaly detection to mitigation. In addition, we discuss how these methods can be used to make cyber systems automated, intelligent, and trustworthy as necessary.
- Our study identifies and summarizes several key challenges and research issues that need to be addressed for further improvement. In addition, we discuss possible next-generation cyber modeling directions within the context of digital twins.

1.2. Article organization

The remainder of this article is structured as follows. The background in related technologies, such as digital twins, cybersecurity, and artificial intelligence, is summarized in Section 2, along with the scope of this work and existing literature. Cybersecurity resilience and potential threats and anomalies in the various functional layers of the digital twin are explored in Section 3. Section 4 presents a comprehensive analysis of AI/XAI approaches for cybersecurity modeling along with their taxonomy building. Different real-world usage scopes are presented in the context of cybersecurity modeling in DT in Section 5, along with a discussion of the potential contributions of these techniques. Section 6 provides a list of research problems and prospects that indicate possible directions. Several key points are outlined in Sections 7 and 8 concludes this paper. In addition, Table 1 contains a list of acronyms and their definitions.

2. State-of-the-art

We begin by exploring the background of digital twins (Section 2.1), cybersecurity in digital twins (Section 2.2), and AI-enhanced cybersecurity (Section 2.3). Then, we review the related surveys within the scope of our study (Section 2.4) to identify the study gap.

2.1. Digital twin

A “Digital Twin” is a digital representation of a physical product, system, or process that typically serves as its virtually identical digital counterpart to simulate and analyze its behavior and performance. DT can be characterized with three main spaces [2] -

- *Physical space:* It includes operational technologies (OTs) that are used in real-world settings, including sensors, actuators, and controllers like remote terminal units (RTUs) and programmable logic controllers (PLCs).

Table 1
List of key acronyms.

Acronyms	Meaning
DT	Digital Twin
AI	Artificial Intelligence
XAI	Explainable Artificial intelligence
AIT	Automation, Intelligence, and Trustworthiness
ML	Machine learning
DL	Deep learning
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
AE	Autoencoder
DBN	Deep Belief Network
KDD	Knowledge Discovery from Data
MLP	Multi-layer Perceptron
NLP	Natural Language Processing
LLM	Large Language Model
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
CIA	Confidentiality, Integrity and Availability
CPS	Cyber-Physical Systems
DDoS	Distributed Denial-of-service
DoS	Denial-of-service
IDS	Intrusion Detection System
IoT	Internet-of-Things
QoS	Quality-of-Services
SIEM	Security Information and Event Management
SOC	Security Operation Centre
SOAR	Security, Orchestration, Automation, Response

- *Digital space:* To represent physical assets using digital assets, it mimics the states or situations, circumstances, and configurations while making decisions about the physical space.
- *Communication space:* It bridges the physical and digital worlds, allowing the DT to control information flows and production processes.

With advancements in technologies like AI, machine learning, and data analytics, digital twins are becoming more advanced and are expected to play a significant role in driving digital transformation across various industries in the future [17,18]. DTs can be used in various industries and critical infrastructures, such as manufacturing, healthcare, business, transportation, energy, water, defense, smart cities, and so on to gain insights, optimize operations, and make data-driven decisions [1,2,19]. Data from the physical object is collected in real time and used to update the digital twin, which in turn provides insights, analytics, and visualization to better understand and manage the physical counterpart. The relevant terms "Digital Model" and "Digital Shadow" can be distinguished according to the data flow and interaction between the physical and digital entities, as shown in Fig. 3. A digital model does not include any automated data flows between the digital and physical worlds, is more or less static (unless manually updated), and exists in isolation. On the other hand, a digital twin has a connection between the digital and physical worlds with fully integrated data exchanges in both directions. A digital shadow lies in between with automated data flowing from the physical world to the shadow, but not from the shadow to the physical world. Overall, digital models, digital shadows, and digital twins can be differentiated based on their data flow architectures as well as their intended uses. Digital shadows are more capable than digital models but less capable than digital twins.

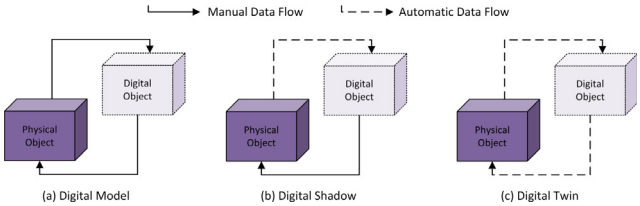

Fig. 3. An understanding of digital twin comparing with digital model and digital shadow by taking into account the data flow between physical and digital object.

Overall, DT technology involves creating a virtual replica of a physical system or device to improve efficiency, reduce costs, enhance performance, and enable innovation by allowing organizations to better understand and manage their physical assets and processes in a virtual environment. In terms of cyber threats and security, DT technology raises new challenges and opportunities in real-world application areas, discussed in the following.

2.2. Cybersecurity and digital twin

Cybersecurity typically involves numerous measures and technologies that ensure the confidentiality, integrity, and availability of data, and safeguard digital assets [31]. Cybercriminals have become increasingly sophisticated in the real world of cyberspace, and the evolution of computer crime towards the use of ICT and AI technologies can be summarized as cybercrime, computer crime, and AI crime [10,11,32].

Digital twins and cybersecurity can be characterized into two different but related perspectives such as "DT for cybersecurity" which is about DT technology as a tool or solution to enhance cybersecurity resilience in a system or organization [33]. This means leveraging the capabilities of DT technology to simulate, model, and analyze potential cyber threats, vulnerabilities, and attack scenarios to proactively identify and mitigate cybersecurity risks. Another perspective could be "Cybersecurity for DT", which is about implementing cybersecurity measures to protect DT environments from potential cyber threats [34]. This means implementing appropriate cybersecurity controls, practices, and technologies to safeguard the digital twin environments from potential cyber threats, such as unauthorized access, data breaches, or other malicious activities. To mitigate today's cyber threats, industries, and businesses need to move towards a more proactive and predictive approach, which can be achieved by using digital twins. As mentioned earlier, DT has the such capability to provide organizations with better decision-making insights to enhance the cybersecurity resilience of their infrastructure. However, it is also crucial to ensure the security of DT itself

Table 2
Previous survey comparison by taking into account ten key aspects relevant to this paper.

Papers	Aspects										
	CyberAIT Aspects	Exploring DT Threats	AI-based Cybersecurity	ML/DL-based Cyberlearning	XAI Methods	Explainable Cyber Decisions	AI/XAI Taxonomy	AI/XAI based Cyber Usage	Challenges and Research Issues	Future Directions and Prospects	Remarks
Alcaraz et al. [2], 2022	x	✓*	x	x	x	x	x	x	*	✓	An extensive study on security threats in digital twin
Rathore et al. [19], 2021	x	x	x	x	x	x	*	x	*	✓	Exploring AI, ML, big data in digital twin
Kaloudi et al. [10], 2020	x	x	✓	✓	x	x	x	✓	*	*	Presented AI-based cyber threat landscape
Hu et al. [20], 2022	x	x	✓	✓	x	x	x	✓	*	*	Presented AI-threats and countermeasures
Kaur et al. [21], 2022	x	x	x	x	✓	x	✓	x	✓	✓	Presented a review on trustworthy AI
Guidotti et al. [14], 2018	x	x	x	x	✓	x	x	x	*	*	Presented a brief summary of XAI for black box systems
Kuzlu et al. [22], 2021	x	x	✓	✓	*	x	x	✓	*	*	Explored the role of AI in the IoT cybersecurity
Samtani et al. [23], 2020	*	x	✓	✓	*	x	x	✓	*	*	Offered a multi-disciplinary AI for Cybersecurity
Alazab et al. [24], 2021	*	x	*	*	x	x	x	*	✓	✓	An extensive study on federated learning for cybersecurity
Dwivedi et al. [15], 2021	x	x	x	x	✓*	x	✓	*	*	*	An extensive study on XAI techniques and tools
Arrieta et al. [25], 2020	x	x	x	x	✓	x	✓	x	✓	✓	Presented an extensive study on XAI
Capuano et al. [16], 2022	x	x	*	✓	✓	✓	*	✓	*	*	Presented an extensive study on XAI in cybersecurity
Seale et al. [26], 2022	x	x	*	✓	✓	✓	✓	*	*	*	Exploring X-IDS methods in cybersecurity
Rawal et al. [27], 2022	x	x	x	x	✓	x	✓	x	✓	✓	Presented recent advances in trustworthy XAI
Charmet et al. [28], 2022	x	x	*	✓	*	*	*	x	*	✓	A study on XAI in cybersecurity
Ahmed et al. [29], 2022	x	x	✓	✓	✓	x	*	x	✓	✓	Presented a study form AI to XAI in Industry 4.0
Ibrahim et al. [13], 2023	x	x	x	x	✓	x	✓	x	✓	✓	Presented Explainable CNN focused XAI
Saeed et al. [30], 2023	x	x	x	x	✓	x	✓	x	✓	✓	Presented a systematic meta-survey on XAI
This paper (Sarker et al.)	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	✓*	An extensive study on AI/XAI for cybersecurity focusing diverse cyber issues in DT, taxonomies, multi-dimensional cyber usage scopes, challenges research directions from the perspective of Cyber AIT (Automation-Intelligence-Trustworthiness).

Symbol Used: High Coverage (✓*), Mid Coverage (✓), Low Coverage (*) and No Coverage (x).

to protect against cyber threats that could compromise the integrity and effectiveness of the DT system and the physical counterpart it represents, discussed briefly in Section 3.

2.3. AI-enhanced cybersecurity in digital twin

The widespread usage of DTs raises several concerns about cybersecurity that need to be addressed to utilize the full potential of DT in our real-world application areas. AI-enhanced cybersecurity plays a critical role in protecting digital twin environments from cyber threats, and making the systems automated and intelligent. In Section 1, we defined these key terms automation, intelligence as well as trustworthiness. AI is typically involved with training machines to think in such a way that requires intelligence to enable machines to carry out certain jobs. AI systems can be of different types like analytical, functional, interactive, textual, and visual depending on the nature of the problem and data [9,18]. Similar to deep learning taxonomy, presented by Sarker et al. [35], AI can also be categorized into generative, discriminative, and hybrid by taking into account the modeling type and outcome, defined below -

- Generative AI — typically focuses on modeling the underlying data distribution and generating new content or data accordingly, *e.g.*, to generate realistic synthetic data, simulating attacks, etc.
- Discriminative AI — typically focuses on finding a decision boundary given a set of input features and to classify or predict the output for new data points accordingly, *e.g.*, intrusion or anomaly detection, identifying suspicious activities, etc.
- Hybrid AI — typically focuses on solving more complex problems that require both generative and discriminative approaches, *e.g.*, incident response by identifying the source and nature of a cyber attack and generating responses to mitigate.

The choice of which approach to use depends on the specific task and the nature of the data involved. For example, a hybrid cyber model could use a generative model for anomalous patterns of network traffic, and a discriminative model to classify the type of attack that is being carried out. Machine learning techniques including deep learning are broadly used in such AI modeling processes which typically allow computers to automatically learn from data and to solve a particular cyber issue [36,37]. Thus, the cybersecurity aspects of a DT can be enhanced and automated using AI. In our context of study the term "Explainable" focuses on high-performance AI models with the capability of human-understandable decision-making. Researchers often use the terms "explainability", "interpretability", "trustworthy", "reliability", and so on interchangeably in different contexts. This can also be considered as the foundation of "Responsible AI" aiming to ensure that AI technologies are developed and used in ways that benefit individuals, communities, and society as a whole. Hence, we broadly take into account these phrases as a problem under the category of explainable AI. Various types of AI techniques for cybersecurity modeling and their explainable

capabilities within the context of our study are discussed in Section 4.

2.4. Related surveys and study scope

The development trend of the DT research is growing rapidly highlighting the formation stage, incubation state, and growth stage [38]. A comparison of related surveys is presented in Table 2. We base our position on ten relevant key aspects, as shown in Table 2. The scope of the study and the contributions of our work are then highlighted by a comparison with earlier surveys from the viewpoint of these key aspects. Three main attributes are used in this survey, *i.e.*, AI, cybersecurity, and digital twins. Thus, we mainly use a combination of the search keywords "Cybersecurity", "Artificial Intelligence", "AI", "Explainable AI", "Machine Learning", "Data Science", "Digital Twin" etc. while searching relevant papers. Peer-reviewed scientific journals, conferences, and books published between 2010 and 2024 are taken into consideration. In terms of databases, we consider several popular repositories such as "Google Scholar", "Science Direct", "Springer Nature", "Scopus", "ACM", and "IEEE Explore". Through our review of the articles' abstracts, introductions, discussions, and conclusions, we assess the paper's relevance to our study in this paper. Eventually, we have listed 142 papers to support our study, 18 of which are survey papers comparing our study listed in Table 2. Some surveys concentrate on DT but have no exploration to do with AI-enhanced cybersecurity [1,2,17]. For instance, Barricelli et al. [1] present a DT survey focusing on definitions, characteristics, applications, and design implications. Alcaraz et al. [2] analyze the current state of the DT paradigm and their associated security threats. Wagg et al. [17] present the state-of-the-art prospects of DT for engineering dynamics applications. Similar to this, some surveys concentrate on AI but ignore the potential of trustworthiness in decision-making [12,19]. For instance, Rathore et al. [19] present a review of the role of AI, machine learning, and big data in digital twinning. Some surveys like [14,16,29] concentrate on XAI methods, particularly focusing on black-box modeling. In addition to black-box modeling, some other works have been conducted with additional techniques in the context of cybersecurity [39]. In general, the choice between black box and white box models depends on the application's specific requirements, including transparency, interpretability, accuracy, and regulatory compliance. A list of related works and their key objectives can be found in Table 2. The lack of a comprehensive study of AI/XAI-based modeling in cybersecurity and DT motivates us to conduct this survey. Thus, we begin with a variety of security concerns in DT in our paper. Following that, we provide a thorough analysis of several AI/XAI techniques for cybersecurity modeling using a taxonomy that includes their explicable features. This can help analysts and security specialists identify potential threats and anomalies, understand how the system functions, and ultimately determine the best path to take. We also go over the potential applications of AI/XAI-based techniques in various cyber domains within the context of DT. We conclude by highlighting the research challenges that have been identified and suggesting possible study avenues for further cyber research and development in DT. We cover all ten key aspects, including CyberAIT, which makes our survey unique compared with other studies in this emerging field.

3. Cybersecurity resilience and threats in digital twin

Digital twins can be used to improve cybersecurity resilience, but it is also important to consider the potential security threats that may arise in digital twin systems [34]. To explore this, we first formulate two questions "How does digital twin enhance cybersecurity resilience using AI?" and "What are the potential security threats of digital twin, and how AI can help to address these issues?" In this section, we answer these two questions and discuss the need and potential of AI to address this.

3.1. Enhancing cyber resilience

Digital twin technology can enhance cybersecurity resilience by providing a virtual representation of a physical system or object, which can be used for testing, monitoring, and analysis of cybersecurity threats [33]. Hence we summarize how digital twin technology can enhance cybersecurity resilience using AI technologies:

- To simulate cyber attacks and test the security of a system before it is deployed. AI-based algorithms can be used to create realistic attack scenarios and test the effectiveness of cybersecurity measures. By analyzing large amounts of data in real-time, AI algorithms can identify anomalies and potential security threats more quickly and accurately than human analysts [36].
- To monitor physical systems in real-time, providing early warning of cyber attacks or anomalies. AI algorithms can be used to analyze data from sensors and identify patterns that may indicate a security breach.
- To analyze historical data and predict future security threats using AI technologies through identifying patterns and trends. This can enable organizations to take proactive measures to prevent cyber attacks before they occur.
- To automate incident response, enabling digital twins to quickly and effectively respond to security threats. For example, AI-powered digital twins can automatically isolate compromised devices, block malicious traffic, and implement security protocols to prevent further damage.
- To simulate and analyze user behavior, providing data to train AI models to detect anomalous user behavior that may indicate a security threat. This can help organizations detect insider threats and prevent data breaches.
- To create intelligent access control systems that can authenticate and authorize users based on their behavior, location, or other factors, to prevent unauthorized access. AI algorithms can be used to detect anomalies in user behavior and alert security personnel.

Overall, AI can play a crucial role in strengthening cybersecurity resilience in digital twins. By leveraging machine learning algorithms and other AI technologies, discussed in Section 4, organizations can improve threat detection, vulnerability assessment, incident response, predictive maintenance, and access control, ensuring the integrity and safety of their digital twin environments.

3.2. Security threats in the digital twin

Digital twin technology provides many benefits, such as enhanced system monitoring, predictive maintenance, and optimization, but it also poses significant security risks or threats [2,34]. According to [40], a security threat can be defined as "a set of circumstances that has the potential to cause loss or harm". In this section, we explore the possible threats of digital twins considering the layer-by-layer architecture of a digital twin. Motivated by our earlier paper Sarker et al. [41] and Alcaraz et al. [2], we take into account the 4-layered architecture of a digital twin and their associated security issues in the following.

3.2.1. Physical layer

The physical layer of a digital twin refers to the hardware and infrastructure that make up the physical environment of the digital twin. This includes sensors, actuators, communication networks, and other physical devices that are typically used to collect and transmit data to the digital twin [1]. Some common security threats in the physical layer of digital twins include physical attacks that can include theft, tampering, or sabotage of the physical system or its components [2,42]. For example, an attacker could tamper with sensors or other critical components of the system, causing inaccurate data to be generated and potentially leading to erroneous decisions. Similarly, malfunctioning hardware components can cause inaccuracies in sensor readings or other data, leading to erroneous decisions. Environmental factors, such as temperature, humidity, and electromagnetic interference, can also pose a threat to physical systems. In addition, malicious code can be introduced into the physical system through infected software, firmware, or hardware. Attackers may use malicious code to gain unauthorized access to the system, steal data, or disrupt its operations. Thus, security threats to the physical layer of a digital twin can have serious consequences, as they can result in physical damage, loss of data, and disruption of operations. AI-based solutions, discussed in Section 4 can be used to detect and prevent such security threats by monitoring and analyzing sensor data, diagnosing hardware failures, modeling the effects of environmental factors, and detecting and preventing cyber attacks on the physical system. For example, anomaly detection using machine learning algorithms [43,44] can be used to identify unusual behavior in control systems or unexpected changes in sensor readings that deviate from the norm as potential attacks.

3.2.2. Data and communication layer

The data and communication layer of a digital twin typically facilitates the exchange of data between the physical system and the digital twin, as well as between different components of the digital twin. Security threats in this layer can arise from various sources, including data breaches, cyber attacks on communication channels, and unauthorized access to the digital twin. For example, data breaches can occur when sensitive data is accessed or stolen by unauthorized parties. Malware and viruses can infect the data and communication layer of a digital twin, compromising the integrity and availability of data. Similarly, insider threats can occur when authorized users misuse their access to the digital twin, either intentionally or unintentionally. AI-based solutions, discussed in Section 4 can be used to detect and prevent security threats in the data and communication layer of a digital twin by monitoring network traffic, identifying suspicious activity, and flagging potential threats. For example, machine learning algorithms [43,44] can be trained to detect unusual patterns of data access or data transfer by authorized users or to detect unusual patterns of network traffic that may indicate malware or virus activity.

3.2.3. Digital and analytical layer

The digital and analytical layer of a digital twin typically monitors physical twin behavior, performs data analysis, and generates insights and recommendations based on the data received from the physical counterpart. Security threats in this layer can arise from various sources, including data tampering, model poisoning, and algorithmic bias. For example, data tampering can occur when data is modified or deleted by unauthorized users, resulting in inaccurate analysis and faulty recommendations. Similarly, model poisoning can occur when attackers manipulate the training data or the machine learning algorithms used to generate insights and recommendations, resulting in biased or inaccurate results [45]. Algorithmic bias could be another issue that can occur when AI algorithms used to generate insights and recommendations are biased against certain groups or individuals, resulting in unfair or discriminatory outcomes. In addition, adversarial attacks can occur when attackers attempt to manipulate the input data to mislead or confuse the machine learning algorithms, resulting in inaccurate or misleading results [46,47]. AI-based solutions, discussed in Section 4 can be used to detect and prevent security threats in the digital and analytical layer of a digital twin by monitoring data integrity and data access, detecting anomalies in the input data and the output results, and preventing algorithmic bias and adversarial attacks. For example, machine learning algorithms [43,44] can be trained to identify patterns in sensor data and adjust the parameters of the digital twin's models to improve accuracy.

3.2.4. User and application layer

The Application Layer of a digital twin is responsible for providing the user interface and functionality for interacting with the system. Security threats in this layer can arise from various sources, including phishing attacks, unauthorized access to user accounts, and vulnerabilities in the user interface. For example, insecure user accounts can be compromised

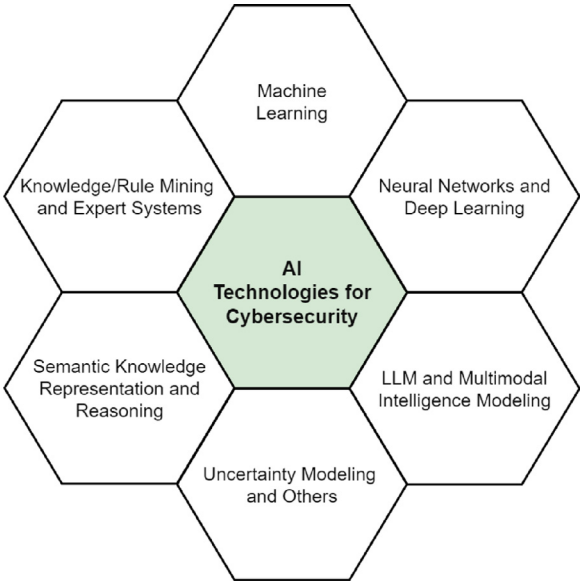

Fig. 4. Major AI technologies for automation and intelligent decision-making in the context of cybersecurity.

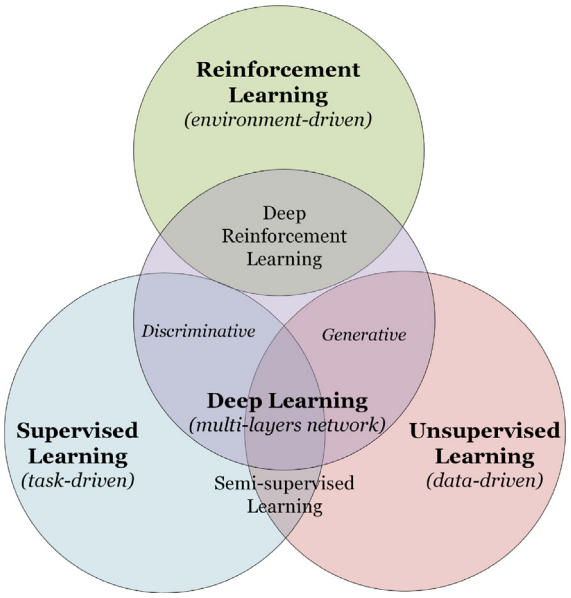

Fig. 5. Machine learning taxonomy highlighting major learning categories used for self-learning cyber automation and intelligence.

through weak passwords or social engineering attacks, allowing unauthorized access to the digital twin. Phishing attacks are a common form of social engineering that trick users into providing sensitive information, such as login credentials or financial information [48,49]. Similarly, vulnerabilities in the user interface can occur when attackers exploit weaknesses in the design or implementation of the user interface. Thus, security threats to the application layer of a digital twin can compromise the confidentiality, integrity, and availability of data, as well as the functionality of the system. AI-based solutions, discussed in Section 4 can be used to detect and prevent security threats in the user and application layer of a digital twin by monitoring user behavior, detecting anomalies in application behavior, and preventing unauthorized access to resources. For example, machine learning algorithms [43,44] can be trained to detect discrepancies between the digital twin's outputs and the actual physical system's behavior or to identify unusual patterns of application usage.

Overall, by integrating AI-based solutions, as discussed briefly in Section 4 into the digital twin systems, cybersecurity can be improved, and the accuracy and reliability of the digital twin can be enhanced.

4. AI/XAI methods and taxonomy

In this section, we explore multi-aspects AI/XAI methods that are useful for cybersecurity modeling in DT and build a taxonomy accordingly, as shown in Fig. 6. To achieve this goal, we first summarize and discuss the potentiality of diverse AI methods (Section 4.1) and then discuss their explainable capabilities (Section 4.2) from different perspectives. Several popular techniques and their potential applications in cyberspace are outlined in Table 3.

4.1. AI methods for cybersecurity modeling

In cybersecurity modeling, artificial intelligence offers advanced techniques for threat detection, risk assessment, anomaly detection, and incident response. To comprehend the potential of diverse AI methods, we first classify them into six key categories based on their working principles that can be used for cybersecurity models, as depicted in Fig. 4. The following subsections discuss these methods, emphasizing their potential to make cybersecurity systems automated and intelligent.

4.1.1. Machine learning

The rapid growth of data generated by digital systems and the complexity of threats are making traditional methods of detecting and preventing cyberattacks less effective. ML, a core component of AI, has the potential to automate the process of detecting and responding to threats, as well as provide more effective and efficient cybersecurity solutions, including intrusion detection, spam detection, malware detection, fraud detection, and user behavior analytics [44,64,65]. A key advantage of machine learning in cybersecurity is its ability to analyze vast amounts of data gathered from network traffic, user behavior, and system logs to identify patterns and anomalies that may indicate a potential security threat, which cannot be done manually by humans. In a broader perspective, learning can be supervised (task-driven), unsupervised (data-driven), and reinforcement (environment-driven) as shown in Fig. 5. Semi-supervised could be another type combining supervised and unsupervised learning. These can be used in various cyber application areas depending on the problem nature and availability of data [66]. For instance, Cui et al. [67] demonstrate ML-based methods, *e.g.*, k-means clustering, and naive Bayes classification with Monte Carlo

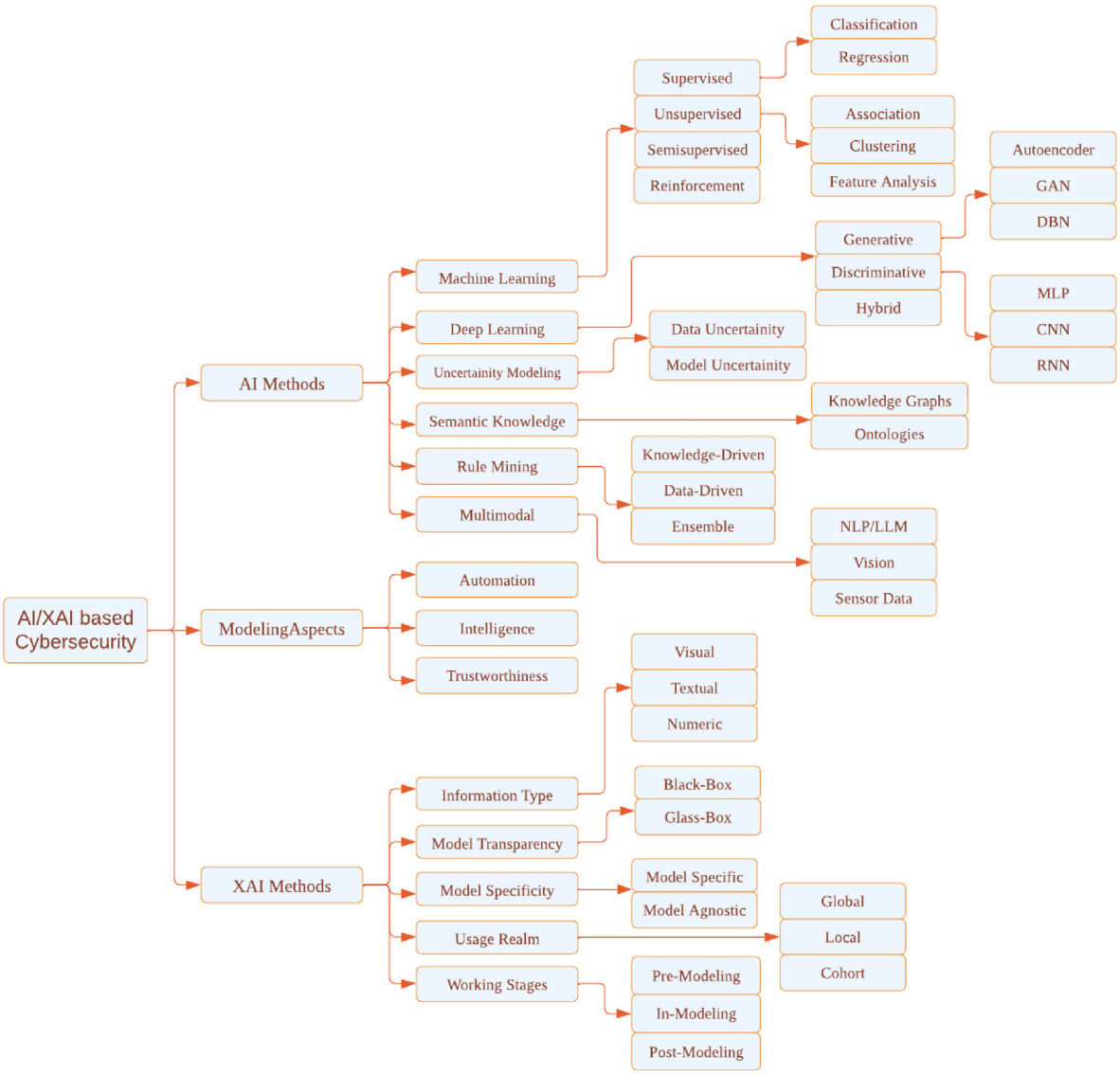

Fig. 6. A taxonomy of AI/XAI based methods for cybersecurity modeling.

simulation, to detect anomalies for load forecasting under cyberattacks. A decision tree-based intelligent intrusion detection system has been discussed in [68]. Heartfield et al. [69] present a self-configurable cyber–physical intrusion detection system for smart homes using reinforcement learning. A tree-based model has been presented in [70] for online diagnosis services and in [71] for stealthy cyber-attack detection in smart grid networks. Similarly, a variety of use cases using machine learning techniques and their potential for security modeling have been summarized in [43]. Based on the data and modeling variations, different categories of ML modeling are used to solve cyber issues. For instance, Alazab et al. [24] highlight the potentiality of federated learning in cybersecurity, which enables different devices to learn a collaborative ML model. An active learning-based XGBoost model for cyber–physical system against generic false data injection attacks has been presented in [65]. However, the key challenge is to ensure the availability of high-quality and labeled data for training ML models, especially considering the complexity and diversity of cyber threats. In the context of adversarial machine learning, robust models are designed to withstand adversarial attacks such as intrusion detection system evasions or data poisoning attacks [31]. Overall, ML methods and their variations could be one of the most promising tools for future-generation cybersecurity systems, particularly when focusing on self-learning automation and intelligent decision-making in DT environments.

4.1.2. Neural networks and deep learning

Deep learning (DL), is a subset of a larger family of ML that uses multiple layers to gradually extract higher-level features from the raw input, and can also be leveraged for cybersecurity modeling in DT environments. DL specifically neural networks with multiple hidden layers, can learn complex patterns and representations from large amounts of data, to identify patterns and make predictions, which can be valuable for cybersecurity in digital twin systems. DL

Table 3
Summary of Various AI/XAI-based Methods Used in the Context of Cybersecurity Applications.

Broad area	Cyber applications	Methods used	Explainability	Main contributions
EAI/ML [50]	Intrusion Detection	Gini Index, DT	Whitebox	To build a generalized IntrudTree with the top-ranked security features
AI/ML [36]	Anomaly Detection	Pearson correlation, DT, RF, XGB, ANN, etc.	–	Detecting cyber-anomalies and multi-attacks with feature importance
EAI/ML [51]	Intrusion Detection	Classifier, SHAP	Global, Local	A framework to improve the interpretation of IDSs
EAI/ML [52]	Fraud Detection	LR, Autoencoder, NN, LIME, SHAP, etc.	Global	Exploring explainability methods for runtime tradeoffs on supervised and unsupervised models
EAI/ML [53]	Anomaly Detection	One Class SVM, DT, K-Means, Rules	Model-agnostic	To explain the anomalies detected by an unsupervised OCSVM ML model through rules
EAI/ML [54]	Intrusion Detection	DT, Human Expert Rules	Whitebox, Rule-based	To propose a rule-based interpretable and explainable hybrid intrusion detection system
EAI/ML [55]	Malware Detection	LR, DT	Whitebox	To propose a hardware-assisted malware detection framework using explainable machine learning
EAI/DL [56]	Phishing Detection	Faster-RCNN, Transfer Learning	Visual	To design a hybrid deep learning system for phishing identification
EAI/DL [49]	Phishing Threat Intelligence	Attention mechanism, Multimodal	Visual	Designing a multi-modal hierarchical attention model for phishing website detection.
EAI/DL [57]	Malware Detection	MLP, NLP, Semantic Rule, Attention mechanism	Textual, Expert Analysis	Designing a ML-based approach to interpret the core malicious behaviors within apps.
EAI/DL [58]	Malware Family Identification	Deep Learning	Visual	Designing an interpretable deep learning model for mobile malware and family identification.
AI/DL [48]	Phishing detection	K-medoids, DT, Searching, Optimal Features, ANN	–	Designing a NN phishing detection model based on decision tree and optimal feature selection.
AI/DL [59]	IoT Threat detection	NLP, TFIDF, LogTF, CNN	–	Designing combined DL approach to detect the pirated software and malware-infected files across the IoT network
AI/DL [60]	Botnet detection	Fuzzy rules, ANN	–	Designing a fuzzy logic based feature engineering method for botnet classification
EAI/DL [61]	Botnet traffic detection and classification	CNN, SHAP	Model agnostic, Global	Designing a DL model for botnet detection and classification with decision explanation
EAI/ML [62]	Twitter bot detection	Ensemble, LR, CART, MLP, AdaBoost, RF, LIME	Visual	Designing a ensemble ML approach for explainable and multi-class bot detection
EAI/ML [63]	Cyber–physical systems	SOM, ANN, Histograms, U-Matrix, Heat map	Global, Local	Designing an explainable unsupervised machine learning for cyber–physical systems

can be categorized into generative (*e.g.*, GAN), discriminative (*e.g.*, RNN), and hybrid modeling, discussed briefly in Sarker et al. [35] and have many potential security applications. For instance, Lv et al. [72] utilized DL-based methods, *i.e.*, CNN-SVR, to solve the security problems of the cooperative intelligent transportation system in digital twins. Luo et al. [73] discussed various aspects of deep learning-based methods (*i.e.*, DNN, CNN, LSTM) to identify anomalies in cyber–physical systems and ensure the security of CPS. An interpretable deep learning model for mobile malware detection and family identification has been presented in [58]. Danilczyk et al. [74] present a smart grid anomaly detection method using a deep learning (CNN) digital twin that can classify the faults with over 95% accuracy. A deep RNN-based approach for IoT malware threat hunting approach has been presented in [75]. DL techniques in cybersecurity are advantageous because they can learn from large volumes of data and identify complex patterns that traditional approaches may overlook. However, much attention might be needed to ensure the effectiveness of the DL model and avoid potential biases or false positives in DT environments.

4.1.3. Rule mining and expert system modeling

Typically, knowledge mining involves extracting insights, patterns, and relationships from large volumes of data [9]. An important part of knowledge mining is rule mining, which is used to find interesting relationships between variables in large datasets. Using discovered knowledge or rules and expert

systems in cybersecurity modeling allows for enhanced threat detection and mitigation. In rule mining, patterns, and correlations are automatically discovered in vast datasets, which can be used to identify vulnerabilities and malicious behavior. Based on these rules, an expert system can be constructed that emulates the decision-making capabilities of human cybersecurity experts. To discover rules, Sarker et al. [11] explored a taxonomy of diverse methods such as knowledge-based approach, *i.e.*, based on human expertise, data-driven approach, *i.e.*, extracting insights or useful knowledge from data, and their ensembles. Different types of rules can be discovered depending on the data nature and the target cyber solution. For example, association rules [76] can be employed to identify a correlation between certain user behaviors and the likelihood of a security breach. Similarly, classification rules [77] can be used to identify the type of malware that is present on a network and to determine the appropriate action to take to mitigate the threat. In addition, fuzzy rules [78,79] and belief rules [80,81] based modeling can be used to handle uncertainty and imprecision in data. For instance, a fuzzy rule could be used to detect network traffic that is slightly anomalous but not necessarily indicative of a specific attack. Similarly, a belief rule could be used to determine the likelihood that a particular security event is a false positive, based on the level of confidence in the detection algorithm and other relevant factors. By using rule-based modeling, security analysts can detect and respond to threats more effectively and find the root cause of such threats for proactive solutions. Through continuous learning and adaptation, rule mining and expert systems enhance the resilience and security posture against evolving cyber threats. However, a balance between rule complexity and interpretability is crucial to effective cybersecurity decision-making. Thus, much attention is needed while designing innovative algorithms by taking into account essential rule properties such as completeness, non-redundant, conflict-free, generalization, and eventually higher accuracy [11] to solve a particular cybersecurity issue.

4.1.4. LLM and multimodal intelligence modeling

The Large Language Modeling (LLM) approach has significant potential for revolutionizing cybersecurity modeling by harnessing advanced natural language processing (NLP) [9]. The use of NLP typically facilitates the identification of potential vulnerabilities and suspicious activities by extracting valuable insights from unstructured text data, such as security logs, incident reports, and threat intelligence feeds. LLMs enable deeper semantic understanding and contextual reasoning by understanding and generating human-like text. The use of LLMs can allow cybersecurity systems to efficiently sift through vast amounts of unstructured data, identify subtle indicators of malicious activities, and make better-informed decisions in real time. However, much attention is needed to take into account diverse issues, *e.g.*, data poisoning, fine-tuning, and trustworthiness in different phases, such as pre-modeling, in-modeling, and post-modeling summarized by Sarker et al. [82]. To generate a comprehensive understanding of potential threats, multimodal intelligence refers to the ability to process and integrate cyber information from various modalities in DT environments, such as text, images, audio, and sensors, rather than relying solely on one type of data. For example, a cybersecurity model may analyze network traffic data, *e.g.*, text, and detect patterns of suspicious activities in DT. It may also analyze images from security cameras and detect unusual behavior or identify individuals who are not authorized to be on the premises. Similarly, by analyzing both the content of emails and the network traffic associated with those emails, a system can more accurately detect and prevent phishing attempts. A variety of machine learning, statistical, and NLP techniques are used in textual analytics to extract insights and patterns from massive amounts of unstructured text [31]. For instance, NLP techniques (word embedding GloVe + CNN) are used to prioritize vulnerabilities based on their description [83]. Similarly, visual analytics extracts insights from images or visual data. Chai et al. [49] present an explainable multi-modal hierarchical attention model by taking into account both the textual and visual information for developing phishing threat intelligence. Thus, by analyzing data from multiple modalities simultaneously, an AI system can identify patterns that might not be visible with a single modality and improve the accuracy of its predictions and alerts. Although getting access to diverse datasets in DT is a challenging issue, incorporating multimodal intelligence into cybersecurity modeling can lead to more effective and efficient detection and prevention of cyber threats.

4.1.5. Semantic knowledge representation and reasoning

By encoding domain-specific knowledge and facilitating intelligent decision-making, semantic knowledge representation and reasoning offer a robust framework for advancing cybersecurity modeling. Using semantic technologies such as ontologies, *i.e.*, formal representations of knowledge within a specific domain, and knowledge graphs, *i.e.*, structured representation of knowledge that captures entities, their attributes, and the relationships between them, enabling rich data integration and analysis, cybersecurity models can capture intricate relationships among threats, vulnerabilities, assets, and defensive measures. These semantic techniques are used in various application areas such as security monitoring [84], malware analysis [85] etc. For instance, a semantic knowledge graph might represent the relationships between different actors, such as threat actors, organizations, and malware families. By representing this information in a structured format, cybersecurity professionals can analyze and reason about potential threats and responses. Wang et al. [86] present a scheme to integrate knowledge reasoning and semantic data for smart factories where the reasoning engine analyzes the ontology model with real-time semantic data. Overall, this structured representation enables sophisticated reasoning capabilities, allowing models to infer complex insights, identify potential attack scenarios, and recommend tailored countermeasures based on contextual understanding. However, much attention is needed to design efficient algorithms and scalable inference mechanisms to detect anomalies, identify patterns, and infer actionable insights

from large-scale semantic knowledge bases. The use of machine learning and knowledge or rule mining methods [11] can augment knowledge graphs through tasks such as entity linking, node classification, relation extraction, recommendation, searching, disambiguating, feature engineering, as well as construction automation, making these applications more useful and effective.

4.1.6. Uncertainty modeling

Due to the inherent complexity and dynamic nature of interconnected systems, cybersecurity modeling within digital twin environments may pose uncertainty issues. These issues arise from a variety of sources, including training data, evolving threat landscapes, and uncertainties in system behavior and interactions. Uncertainty modeling in AI encompasses two primary dimensions: data uncertainty and model uncertainty. Data uncertainty arises from limitations in the quality, quantity, and representativeness of available data, such as noise, bias, missing values, outliers, incompleteness, and variability in real-world phenomena. Model uncertainty, on the other hand, refers to the assumptions and limitations inherent in the algorithms resulting from architectural complexity, parameter estimation, and generalization capability. Data uncertainty can often be addressed through robust preprocessing, data augmentation, or statistical methods, but model uncertainty often requires more sophisticated approaches. A variety of techniques are employed to address these different types of uncertainty, including probabilistic graphical models, fuzzy logic, belief functions, Bayesian inference, or Monte Carlo methods [87,88]. Models using probabilistic graphical representations, such as Bayesian Networks, provide a structured framework to represent and reason about uncertainties, while fuzzy logic allows for flexibility in working with imprecise or qualitative data. Conversely, model uncertainty can be mitigated through techniques like ensemble learning, dropout regularization, model calibration, Bayesian model averaging, and sensitivity analysis [89], which assess how robust the predictions of an AI model are under various assumptions and parameters. To ensure robustness and trustworthiness in AI systems, it is crucial to balance these two aspects of uncertainty modeling.

4.1.7. Others

In addition to the above key categories of AI methods, several other techniques are also useful in the context of cybersecurity modeling. For instance, information fusion, which is 'the study of efficient methods for automatically or semi-automatically transforming information from different sources and points in time into a representation that provides effective support for human or automated decision-making' [90]. Data can be generated from a variety of sources, such as machines, physical environments, virtual spaces, and historical databases, in digital twin systems [38]. By combining this data, analysts can identify potential threats more effectively, such as fusion-based malware detection [91], intrusion detection [92], etc. Data generation, modeling, cleaning, clustering, dimensionality reduction as well as advanced mining techniques can be included in this process [38,93]. The use of feature engineering to select or create relevant features from raw data could improve the performance of AI models. For optimizing security parameters, inventing intrusion detection rules, and analyzing malware, genetic algorithms could be useful. In addition, hybrid intelligence, which combines different AI techniques and methodologies, can be used to solve a range of problems. The combination of machine learning algorithms with expert systems or rule-based systems is one example of hybrid intelligence in cybersecurity modeling [31]. Machine learning algorithms can identify patterns and anomalies that may not be immediately apparent to human analysts, while expert systems or rule-based systems can incorporate domain-specific knowledge and rules to identify potential threats that might not be identified by machine learning algorithms alone. Another example could be the combination of supervised and unsupervised learning techniques [66]. Similarly, different methods such as LLM, semantic knowledge, visual analytics, and machine or deep learning can be integrated to produce an output depending on available data and target solution. For instance, Garrido et al. [84] present machine learning-based knowledge graphs for security monitoring. Piplai et al. [85] use fusion, NLP (named entity recognition, word2vec, TF-IDF score, etc.), and neural networks to create their knowledge graph for malware analysis utilizing action reports. Qaisar et al. [91] present a multimodal information fusion for Android malware detection using lazy learning. Overall, hybrid intelligence could be a valuable approach that can enhance the effectiveness of cybersecurity modeling by leveraging the strengths of different AI techniques and methodologies.

In summary, these AI methods, discussed above are often used in combination to tackle real-world problems, and the choice of technique depends on the specific task and the nature of the available data. The incorporation of these AI methods with their explainability analysis into cybersecurity modeling can enhance the effectiveness of organizations in detecting, mitigating, and responding to cyber threats.

4.2. XAI methods for cybersecurity modeling

In this section, we discuss XAI methods from different perspectives highlighting the explainable capabilities of AI methods discussed earlier. To comprehend XAI methods for cybersecurity modeling, we classify them into five categories, discussed in the following subsections.

4.2.1. Explainability based on model transparency

This category of interpretability is based on how the model is transparent in terms of its architecture such as glass or white-box and black-box modeling, depending on the specific needs of the application. White-box models are typically transparent models, such as decision trees or rule-based systems explored by Sarker et al. [11], *i.e.*, the internal structure and decision-making process of the models are transparent and interpretable. These models can be particularly useful in cybersecurity for explaining how specific security policies or rules are being enforced. For instance, Sarker et al. [50] present

a tree-based machine learning model for intrusion detection, where the anomaly detection rules are generated by traversing from root node to leaf which is human-understandable. The transparency of white-box models makes them ideal for use cases where the ability to understand and explain decision-making is essential. However, the problem with white-box modeling is that it may not be employed to work with complex dependencies that have a lot of parameters. Thus, black-box models, such as deep neural networks [35] may include billions of parameters, and can be used to detect patterns and anomalies in large datasets, such as network traffic or user behavior logs. For instance, Luo et al. [73] presented a survey of deep neural network learning-based anomaly detection in cyber–physical systems. However, these models can be difficult to interpret and understand, making it challenging to determine why a particular decision was made. Decisions made in cyberspace without clear justifications are generally quite impractical due to the trustworthiness issues [9]. To make black-box models more transparent, XAI techniques such as sensitivity analysis, feature importance, and model visualization can be used to reveal how the model is making its decisions. For example, feature importance [36] can show which network traffic features are having the most influence on the model's output, while model visualization can provide a graphical representation of the decision-making process. It might be worthwhile to trade off white-box and black-box modeling according to the requirements [9]. Overall, the choice between black box and white box models depends on the application's specific requirements, including transparency, interpretability, accuracy, and regulatory compliance, may differ application to application.

4.2.2. Explainability based on model specificity

Depending on how the model is specific or agnostic, this interpretability strategy is taken into account. Model-specific techniques such as decision trees, SVM, XGBoost, linear regression, etc. [66] are designed to provide explanations for a specific machine-learning model. A model-specific XAI technique is useful for understanding and explaining the decisions made by a specific model. Transparency and trust can be increased in the decision-making process of models using these techniques. The model-agnostic approach, on the other hand, provides explanations that apply to any machine learning model, regardless of its architecture or algorithm. Examples of model-agnostic XAI techniques include LIME and SHAP which are employed for fraud detection [52] and Botnet traffic detection [61] respectively. A wide range of machine learning models can benefit from these techniques, which can help identify which features of the input data are most significant for a given decision. XAI with model-agnostic techniques is useful when the goal is to understand and explain decisions made by any machine learning model, rather than a specific model. Model specific approaches depend on a certain model structure, *e.g.*, a specific architecture of CNN, whereas model agnostic techniques function with any type of ML model [94]. Flexibility is a key advantage of model-agnostic interpretation techniques over model-specific ones. However, these methods are typically less accurate as they simply use the input and output to explain the behavior of the models while model-specific approaches rely on the characteristics of the particular methods or models.

4.2.3. Explainability based on information type

This category of interpretability techniques is dependent on the form in which explanation data is presented. It might involve visual explanation techniques, such as heatmaps, charts, graphs, etc., that generate images or plots to illustrate the model's decisions. In particular, dimensionality reduction, clustering, classification, and regression analysis play a significant role in the interpretation of the machine learning algorithm. For instance, Wickramasinghe et al. [63] use visualization methods like Histograms, Heat Maps, and U-Matrix (Unified Distance Matrix) to visualize how the feature values change across clusters for their ML-based cyber–physical systems. Szafron et al. [95] visualized the classifier decisions and the supporting data for these decisions using a straightforward graphical explanation to explain the naïve Bayesian, linear support vector machine, and logistic regression classification process. Textual explanation techniques generate natural language text to interpret the decisions [96]. For instance, Wu et al. [57] present a method that generates an understandable natural language description to interpret the malicious behaviors of Android apps. Mathematical explanations or numerical scores could be another format for providing more detailed explanations for the overall findings [97]. For instance, a linear classifier is fitted to the intermediate layers to track the features and assess how well-suited they are for classification.

4.2.4. Explainability based on usage realm

This category of interpretability is dependent on the usage realm such as global, cohort, and local model explainability depending on the characteristics of the model. Global model explainability aims to provide an overall understanding of how an AI model works such as Permutation Importance and SHAP [61,94]. It involves analyzing the model's architecture, training data, and parameters to identify the most relevant features that contribute to the model's outcome. This can help identify the most critical attack vectors and the most important indicators of compromise in the context of cybersecurity. On the other hand, local model explainability focuses on understanding how an AI model makes specific outcomes such as LIME [52,94]. It involves analyzing the model's decision-making process for a particular input, such as a network packet or a log file. This can help identify the specific features that triggered the model's decision, which can be useful for investigating suspicious activities or validating the model's output. A study by Wang et al. [51] demonstrated in their experimental analysis that local explanations explain why models make certain decisions based on inputs, while global explanations present the relationships between feature values and types of attacks extracted from IDSs. Another one is cohort model explainability involves comparing the behavior of an AI model with that of similar models trained on different datasets or with different parameters. By analyzing the differences in behavior,

XAI can identify the specific factors that contribute to the model's performance. For cybersecurity applications, cohort model explainability can help detect anomalies and outliers that may indicate a potential attack or data breach. Overall, these explainable capabilities can provide valuable insights into the behavior of AI models in cybersecurity applications, which can help analysts detect and respond to potential threats more effectively.

4.2.5. Explainability based on working stages

Explainability can be used at every stage of the AI development process and can be divided into three different ways as Pre-Modeling, In-modeling, and Post-modeling stage [9]. Premodel explainability focuses on the data used to train the machine learning model, which aims to ensure that the training data is representative, unbiased, and reliable. By examining the data used to train the model, analysts can identify potential biases, inconsistencies, or errors that may affect the model's performance and robustness. In-model explainability approach involves examining the internal workings of the machine learning model to understand how it makes predictions such as rule-based modeling [98]. In-model explainability techniques can be used to identify the most important features or variables that contribute to the model's output. This can help analysts identify potential vulnerabilities, biases, or errors in the model's decision-making process. Thus this can help identify potential threats or anomalies in network traffic, user behavior, or system logs in the context of cybersecurity. Postmodel explainability involves analyzing the output of the machine learning model to understand how it performs in real-world scenarios. For instance, Langone et al. [99] use a posthoc approach for analyzing anomaly detection and Mehdiyev et al. [100] for predictive analytics. This can be used to validate the model's performance, identify potential errors, or generate insights into the model's behavior, which can help analysts respond more effectively.

4.3. Performance analysis and discussion

Depending on the nature of the problem and the data characteristics, different methods can potentially be used to build AI-based cybersecurity models. Table 3 summarizes methods used in various cyber applications, highlighting their explainability and contributions. While accuracy on unseen test cases is an important metric, other metrics can also be used to assess a model's effectiveness, such as detection rate, false positive rate, false negative rate, error calculation, etc. [11].

For instance, Wang et al. [51] demonstrated their experimental results with accuracy, precision, recall, f1-score, etc. while building their explainable machine learning framework for intrusion detection systems. They achieved 'accuracy = 0.806', 'precision = 0.828', 'recall = 0.806', and 'f1-score = 0.807' for the one-vs-all classifier, and 'accuracy = 0.803', 'precision = 0.828', 'recall = 0.803', and 'f1-score = 0.792' for the multi-class classifier, utilizing NSL-KDD test dataset. For explainability analysis, they use SHAP and combine local and global explanations to improve the interpretation of IDSs. Pan et al. [55] demonstrated their experimental results while hardware-assisted malware detection and localization using explainable machine learning. They achieved their highest results with 'accuracy = 88.9', 'false positive = 5.2', 'false negative = 5.9', and 'f1-score = 0.88' for Decision Tree modeling, and with 'accuracy = 97.7', 'false positive = 0.9', 'false negative = 1.4', and 'f1-score = 0.97' for RNN-LSTM modeling. Ullah et al. [59] demonstrated their experimental results while building their cyber security threats detection model in IoT using a deep learning approach. They achieved results with 'accuracy = 97.46%', and 'f1-score = 97.44%' for their deep convolutional neural network (DCNN) model to detect malicious infections in IoT networks through color image visualization. Chai et al. [49] demonstrated their experimental results while building an explainable multi-modal hierarchical attention model for developing phishing threat intelligence. They achieved results with 'accuracy = 0.97', 'precision = 0.97', 'recall = 0.96', and 'f1-score = 0.97' for their multi-modal hierarchical attention model consisting of URL, webpage text and images.

In general, the performance of the resulting AI model depends on the data characteristics, preprocessing, and intended solution. When analyzing the KDD Cup dataset, Sindhu et al. [101] show that the detection rate is influenced by the features selected. The accuracy of a system can vary depending on the features selected and the categorization, such as binary or multiclass, as described by Sarker et al. [36]. In certain scenarios, multiple methods can be integrated into one approach; therefore, the outcome depends on the integration. Overall, several factors, such as the nature of the problem, the available data characteristics, the computational resources, the interpretability requirements, and eventually the specific project goals, are needed to consider to choose and design an effective AI model.

5. Real-world usage scopes

In this section, we summarize and discuss the potential real-world usage scopes of AI/XAI-based cybersecurity modeling in the digital twin, as shown in Fig. 7, from different perspectives as below.

5.1. Predictive maintenance and proactive solutions

AI-based predictive maintenance can significantly enhance cybersecurity in a digital twin environment by proactively identifying and mitigating potential threats. This typically involves using machine learning and analytics techniques [66, 102] to analyze data from the digital twin and make predictions about potential security threats, vulnerabilities, or risks that may arise in the future, and eventually recommend preventive measures. For example, a DT may recommend software updates, network configuration changes, or other security measures to mitigate the risk of a cyber attack. Thus businesses can proactively uncover security issues before causing any harm with predictive analytics, where machine learning algorithms can be employed. Baryannis et al. [103] present an approach

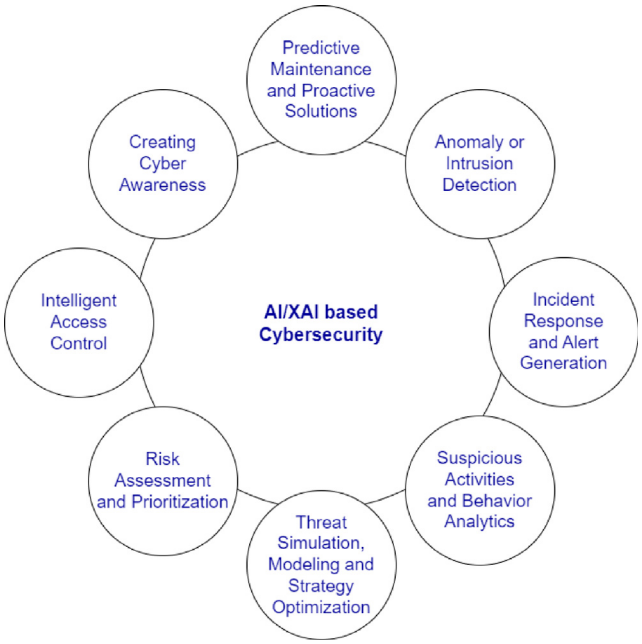

Fig. 7. AI/XAI based potential usage scope in the context of cybersecurity.

to predicting supply chain risks using machine learning algorithms (SVM, Decision Trees). Okutan et al. [104] presented an approach to predicting cyber attacks with Bayesian networks using signals drawn from global events and social media. Fang et al. [105] present a deep learning framework utilizing the bi-directional RNN with LSTM for predicting cyber attack rates, which gives better accuracy than statistical approaches like ARIMA (Autoregressive integrated moving average). Overall, AI-based predictive maintenance can be a valuable tool in the context of cybersecurity within a digital twin, helping organizations to predict and assist in providing proactive solutions to protect critical systems and data from being compromised.

5.2. Intrusions or anomaly detection and classification

Identifying patterns and anomalies in logs and traffic can be accomplished within a digital twin environment by monitoring the behavior of the virtual network. Machine learning algorithms can be used to detect and classify potential threats and anomalies within a digital twin based on observed patterns and behaviors. A machine and deep learning algorithm can be trained on large historical datasets to identify patterns and relationships in the data [36,50]. Data generated from the digital twin can then be analyzed in real time and deviations from the learned patterns detected. Castellani et al. [106] demonstrate real-world anomaly detection using digital twin systems and ML techniques (SVM, Isolation Forest, KNN, PCA, Clustering, CNN-SAE). Balta et al. [107] presented a digital twin-based framework to detect attacks and anomalies for cyber–physical manufacturing systems, where they utilize one-class support vector machines (OSVM) to model normal behavior. Xu et al. [108] present a digital twin-based anomaly detection in cyber–physical systems taking advantage of unlabeled data and continuously learning at runtime, where Generative Adversarial Network is used as the backbone of the framework. Sahingoz et al. [109] demonstrate a phishing detection system using machine learning techniques (DT, Adaboost, RF, SMO, KNN, NB) as well as various features such as NLP-based features, word vectors, and hybrid features. Qiu et al. [110] summarized various uses of deep learning models (FCN, CNN, RNN, DBN, AE, and hybrid) to detect Android malware. Kocher et al. [111] summarized various uses of ML and DL methods for intrusion detection systems. Shafiq et al. [112] present a malicious Bot-IoT traffic detection method in IoT networks using machine learning techniques (DT, NB, RF, SVM). A classification system within the digital twin can also be used to understand the nature of the threats and to prioritize their response accordingly. For example, a classification system might categorize an intrusion as a brute-force attack, a denial-of-service attack, or a malware infection [36,43].

5.3. Suspicious activities and behavior analytics

In a digital twin, users interact with the virtual system in a similar way to how they would interact with the physical system, which makes behavior analytics important in this context. Behavior analytics involves using machine learning algorithms to analyze user behavior and identify patterns or anomalies that could indicate suspicious activity. For instance, Vallathan et al. [113] present a suspicious activity detection approach using deep learning in secure assisted living IoT environments. Similarly, a hybrid deep-learning-based scheme for suspicious flow detection has been presented in [114]. The AI and machine learning algorithms can monitor user activity, including access to systems and data, network traffic, and other behaviors that may indicate a potential threat [43]. This can help organizations detect when a user is attempting to access resources or data that they do not have permission to access, or when a user is accessing resources outside of their normal usage patterns through the power of AI modeling. For example, machine learning can be used to analyze employee activity within a digital twin of an organization's IT infrastructure and detect unusual behavior, such as unauthorized access attempts or data exfiltration. By monitoring user activity, the system can identify potential insider threats and take appropriate actions to prevent data theft or other malicious activities. Overall, by leveraging AI and machine learning algorithms to analyze user behavior, organizations can identify potential security risks and take appropriate action in the digital twin.

5.4. Risk assessment and prioritizing threats

Risk assessment and prioritization using AI in the digital twin has the potential to revolutionize the way organizations manage risks in their systems, processes, and products. Traditional risk assessment and prioritization methods can be time-consuming, and error-prone, and may not take into account all the variables that could affect the likelihood and

impact of a risk. By leveraging machine learning with other AI techniques in the digital twin, organizations can automate the process of risk assessment and prioritization, enabling them to identify potential risks and threats more quickly and accurately [37,115]. This can involve using techniques such as classification, regression, clustering, and natural language processing (NLP) [31,66]. AI algorithms can analyze vast amounts of data generated by the digital twin to identify potential attack vectors and generate new attack scenarios, that might be missed by human analysts. This can help organizations to better understand the potential risks and prioritize their security efforts. For example, several generative AI methods such as Generative Adversarial Networks (GANs) [116], Variational Autoencoders (VAEs) [117], Recurrent Neural Networks (RNNs) [118], Transformer models [119] etc. have the capabilities to process and generate new data and can be used to create new attack scenarios based on the existing vulnerabilities. For instance, Yan et al. [116] presented an architecture that automatically synthesizes DoS attack traces using GANs. Additionally, AI can learn from historical data, enabling it to make more accurate predictions about future risks and their potential impact. Once potential risks are identified, they can be prioritized based on their severity and likelihood of occurrence, which is a crucial factor for decision-making and formulation of mitigation plans [120]. Prioritization can help organizations focus their resources on the most critical risks, enabling them to take proactive measures to prevent or mitigate them.

5.5. Threat simulation, modeling and optimization of security strategies

By simulating attacks or threats and examining the impact on the virtual equivalent of physical twins, digital twins can assist enterprise security [121]. AI-based threat simulation and modeling can be a valuable cybersecurity application in a digital twin environment. Threat simulation involves using AI algorithms to simulate various cyber attack scenarios, while modeling involves creating virtual representations of the digital twin and its components to assess vulnerabilities, evaluate the effectiveness of cybersecurity defenses, and optimize security strategies. For example, AI can simulate different types of cyber attacks, such as ransomware attacks, DDoS attacks, phishing attacks, or insider threats, in the digital twin environment. These simulations can be based on known attack patterns, historical attack data, or even generated using AI-generated adversarial attacks. Generative AI can be used to create realistic simulations of cyber attacks, such as phishing attacks or malware infections, and model the behavior of attackers in response to different security measures. These simulations can provide insights into the potential impact of different attack scenarios, including the propagation of attacks, the exploitation of vulnerabilities, and the potential consequences for the digital twin and its components. Generative AI is particularly useful when real-world data is scarce or difficult to obtain to simulate different scenarios and test the effectiveness of different security measures. This can help organizations assess the vulnerabilities of their digital twin identify potential weaknesses that could be exploited by real-world attacks and develop more effective cybersecurity strategies accordingly. Thus AI can optimize security strategies in the digital twin environment based on the results of threat simulations and vulnerability models.

5.6. Intelligent access control

The digital twin itself, or a secure entity in direct communication with the digital twin, must ensure that access control is implemented to all incoming requests [122]. It includes requests for and exchange of information with third parties as well as exchange of information with other digital twins. AI-based access control systems can dynamically adjust privileges based on risk factors and user behavior. When a user exhibits behavior that is unusual or potentially dangerous, access privileges can be automatically restricted or revoked until the situation is investigated. Heaps et al. [123] presented a dynamic access control policy generation method from user stories information using machine learning (Transformers, CNN, and SVM). Nobi et al. [124] conduct a survey of access control systems using machine learning. AI can be used to analyze user behavior and determine appropriate access permissions based on user roles and access policies. AI can also play a key role in automatically assigning roles and permissions to users based on their job roles, access history, and other factors. In addition, AI can assist in dynamically adjusting access permissions based on the context of a user's request, such as their location, device type, and time of day. Overall, AI-based access control in digital twin can help organizations improve their security posture by providing automated access management processes and reducing the risk of unauthorized access.

5.7. Real-time monitoring, incident response and alert generation

By continuously analyzing operations and network traffic, an AI-based security system can detect unusual behavior that leads to penitential attacks and alert system administrators to mitigate them. This can be done by training the AI system on a dataset of known threats and attacks and then using it to classify new network activity based on these patterns, where machine learning algorithms can play a key role [43]. For instance, Liu et al. [125] propose an intelligent reinforcement learning-based approach that can intelligently learn mitigation policies under various attack scenarios and mitigate DDoS flooding attacks instantly. Alturkistani et al. [126] presented an approach for optimizing cybersecurity incident response decisions in SIEM systems using deep reinforcement learning (deep Q-learning). Hughes et al. [127] presented a deep reinforcement learning-based approach to facilitate the creation of different incident response policies. Bashendy et al. [128] conducted a survey on intrusion response systems for cyber–physical systems focusing on various architectures

and decision-making processes highlighting the recent advances using reinforcement learning algorithms (Q-learning, DQN, SARSA, DDPG, etc.). Overall, AI algorithms have the potential to monitor network traffic, system logs, and other data sources to identify cyber incidents and trigger an appropriate response, *e.g.*, automatically block the attack and notify security personnel.

5.8. Creating cyber awareness to users

To create cyber awareness among users, understanding why it happened is crucial. For this, identifying the root cause of incidents, *i.e.*, diagnostic analytics is needed to discover in a digital twin environment. Diagnostic analytics typically answers the question "Why did it happen" through analyzing past data, to gain insights into why things happened in the past and what actions can be taken to prevent similar issues from occurring in the future. Thus identifying patterns, trends, and correlations in data to determine the root cause of an incident might be helpful. Several data analytics and machine learning techniques such as association analysis, correlation analysis, rule-based analysis, as well as statistical approaches to examine data sets and identify factors [43] could play a key role in this purpose. For instance, Steenwinckel et al. [129] present a method for adaptive anomaly detection and root cause analysis on sensor data streams. This method combines expert knowledge with machine learning techniques. Eckhart [130] et al. presented a method for improving cyber situational awareness in cyber–physical systems through the use of digital twins. Sarker et al. [50] present a machine learning-based intrusion detection model that generates rules from decision trees capable of explaining the cause of anomalies. Thus, creating a cyber-aware culture among users accordingly could be one of the best practices to minimize the risk of cyber-attacks in the digital twin environment.

6. Challenges and future prospects with potential research directions

Based on our extensive study, we identify several challenges and research issues still open in the context of AI/XAI-based cybersecurity modeling in DT environments. In this section, we summarize these challenges that need attention by the researchers and industry experts in this emerging area of study as well as highlight the prospects with potential directions. These are:

- *Data Heterogeneity and Privacy-Aware Self-Learning:* Digital twin environments can be heterogeneous and complex in the real world [131]. Data from multiple sources such as network logs, system logs, user behavior data, or other historical or real-time data, may need to be integrated and analyzed. Both software and hardware-based data collection processes can be used [20]. However, there may be concerns about sharing sensitive data, which makes it difficult to solve using traditional centralized learning techniques. One possible solution could be federated learning [24,132], a decentralized approach that allows organizations to collaborate and share information while keeping the data decentralized and private. For example, federated deep learning can be used for malware detection, where the goal is to identify and block malicious software on multiple devices without compromising the privacy of the users. However, federated learning assumes that data from different parties is similar in distribution and format, which might not be true always in the context of cybersecurity modeling due to network topology, types of attacks, and security policies. Thus cyber researchers need to focus on effectively modeling federated learning or developing privacy-aware efficient techniques with their explainable capabilities to handle heterogeneous data collected from diverse sources in DT.
- *Data Generation and Annotation Issue:* In a digital twin environment, it is essential to have enough data to test and validate the security of the system, which is challenging. Generative AI can be used to generate new data that mimics the characteristics of real-world data [35]. In terms of attacks, Generative AI can be used to create new attack scenarios based on the existing vulnerabilities in the digital twin environment, which can help organizations to better understand the potential risks and prioritize their security efforts. In addition, the lack of labeled data is a significant challenge due to annotation cost, human efforts, and time-consuming issues faced by researchers and practitioners in the field. Therefore, an automated approach with good generalization and decision-making capability is expected to solve this primary issue. Traditional semi-supervised solutions with a certain amount of labeled data might not be effective due to imbalance issues in cyber incident data. Thus the concept of active learning [133] dynamically selecting the most informative samples for labeling or self-supervised learning [134] predicting a target variable from input data to create a supervisory signal could be a possible solution in this context. Another promising research direction in cybersecurity modeling for digital twins could be transfer learning [119]. This involves using pre-trained models on related tasks to improve the model performance with limited data to transfer cyber knowledge from other closely related domains such as network intrusion detection or malware classification. Researchers can investigate the effectiveness of unsupervised learning techniques [43] to cluster data and identify patterns and anomalies without the need for labeled samples as well as data augmentation techniques through generative AI such as generative adversarial networks (GANs) or variational autoencoders (VAEs) [35,135] depending on the nature of target application.
- *Developing Smart Algorithms and Models:* In many cases some traditional algorithms, mentioned in Section 4 might not be effective due to the constantly evolving cyber threats, and new attack vectors are emerging all the time. This dynamic nature can lead to uncertainty in the model's outcome. Scalable and smart algorithms

with their explainability analysis can help effectively and efficiently analyze security logs and events with real-time monitoring and decision-making. Thus cybersecurity researchers need to focus on designing innovative algorithms and models to handle these issues, which could be a promising research area and direction in the context of today's cybersecurity. Another direction could be model optimization and trustworthiness analysis in machine learning [43], deep learning [35], rule mining [11], generative AI [9], LLM [82] or other AI methods based cybersecurity modeling, discussed briefly in Section 4, as it determines the performance and accuracy of the resultant cyber model in DT. Although several techniques such as trail error, grid search, and Bayesian optimization exist, it is important to consider the trade-offs between accuracy and false positives/negatives in the context of cybersecurity modeling. For example, a model that has high accuracy but generates several false positives, might not be practical for real-world cyber usage in a DT environment.

- *Automatic Rule Generation and Security Policies:* This involves data-driven approaches to generate rules and policies that trigger alerts or response actions when deviations from normal behavior are detected. The discovered patterns from the data collected from the DT environment and relevant features can be used to generate rules and policies to detect any deviations from the normal system behavior that could indicate a potential cyber threat. The major advantages of data-driven approaches are creating evidence-based rules according to data patterns, adapting to new updates as well as reducing manual efforts, discussed briefly in Sarker et al. [11], which is very difficult to create and manage rules manually for a large scale DT system. Existing techniques such as association learning [76,98] may not be effective due to producing redundant rules which may lead to inefficient decision-making and computationally expensive. Therefore, developing scalable rule discovery algorithms and eventually making dynamic decisions accordingly could be a significant direction in the context of DT research.
- *Handling Malicious Behavior Changes and Adapting Concept Drift in Cyberspace:* In the real-world scenario in the DT context, concept drift may occur due to changing behavioral patterns over time. Traditional AI algorithms may not be able to adapt to concept drift in cyberspace in DT in real time, which can lead to inaccurate predictions and false alarms. Thus developing adaptive algorithms that can continually learn and adapt to the new types of attacks, could be a promising research direction in this context. Moreover, algorithms that can automatically identify the most relevant features for detecting different types of attacks, could play a key role [136]. In addition, developing incremental learning [137], dynamic updated ensemble learning [138, 139], recent pattern-based mining [140] as well as their hybridization could be a major direction in this context. Investigating how transfer learning [119] can be used, *e.g.*, transferring knowledge from related domains like network traffic analysis, could be another possible domain depending on data availability.
- *Context-Awareness for Adaptive Cybersecurity:* In cybersecurity, context-aware decision-making involves analyzing the context of a potential threat or attack and making a decision based on that analysis. Thus it can be considered a crucial aspect of AI-based cybersecurity as it enables systems to make informed decisions based on the relevant contextual information such as the spatio-temporal, environment, the user, the device, and other factors to decide on how to respond to a security event. Context-aware decision-making is also highly human-interpretable and can help security experts understand and trust the decisions made by the models. For instance, it can allow a system to make a decision based on the current state of the network or the behavior of the user that may vary over time. For instance, context awareness can play a significant role in classifying software vulnerabilities [141], user behavior modeling [98], and so on. Thus developing context-awareness cybersecurity models that can behave according to the current contexts, could be a significant research direction in this domain. Researchers also need to explore new approaches to human–AI collaboration, where AI systems can work together with human experts to make context-aware decisions that are more effective than either system could achieve alone.
- *Enhancing Semantic Knowledge with Extracted Cyber Insights:* The knowledge-driven system such as cybersecurity knowledge graph or ontology-based knowledge representation can represent the complex knowledge of heterogeneous systems with the semantic capabilities [142]. However, it would not be able to detect anomalies since it fails to handle large amounts of raw data and human experts lack the knowledge to manually confirm which patterns in the raw data might indicate potential threats [129]. Thus a hybridization of semantic knowledge representation techniques with machine-learning and knowledge or rule mining methods [11] could be useful in terms of automation, scalability, performance and interpretability in the context of large scale cybersecurity systems. By analyzing data from security logs, incident reports, threat intelligence feeds, and other sources, AI and machine learning algorithms [66] can automatically identify patterns, correlations, and anomalies that can be used to construct a comprehensive and dynamic knowledge graph. For instance, NLP can be used to extract information from unstructured data sources such as text-based logs and reports, and RL can be used to optimize the graph structure based on feedback from SOC analysts. Thus designing ML-enhanced dynamic knowledge graphs that can help security teams better understand their network and quickly respond to potential threats of an organization, could be a significant research direction.

- *AI Model Interpretability and Trustworthiness:* Ensuring the interpretability and trustworthiness of AI models is paramount for effective cybersecurity modeling within digital twin systems. As digital twin systems become increasingly complex, understanding the decisions that AI algorithms make becomes increasingly important for cyber defense strategies. As an example, LLM has strong computing capabilities, but its black-box nature makes it hard to explain the outcome [82]. Interpretability thus plays an important role in allowing stakeholders to validate and trust AI-driven decisions. Achieving trustworthiness involves not only verifying the accuracy of AI predictions but also assessing their resilience to potential threats and vulnerabilities. In addition, ethical and regulatory considerations must be addressed to ensure a responsible deployment of AI in digital twins. Developing techniques to explain AI outputs, quantifying uncertainties, and integrating human feedback is essential for accomplishing these goals. By prioritizing research efforts towards enhancing interpretability and trustworthiness, cybersecurity modeling can leverage AI capabilities while maintaining transparency, reliability, and adherence to ethical standards, thus strengthening the resilience of digital twin systems against evolving cyber threats.
- *Adversarial Attacks:* In the context of digital twin environments, adversarial attacks pose a significant threat to the integrity and reliability of interconnected systems. By carefully crafted perturbations into input data, these attacks exploit vulnerabilities in AI-driven models, which leads to incorrect or compromised decisions [45, 46]. As digital twins become more interconnected and data exchange occurs more frequently, adversarial attacks are more likely to disrupt critical operations and compromise sensitive data. To address this issue, robust defense mechanisms are needed, including adversarial training, anomaly detection, and model verification techniques adapted to the unique characteristics of digital twin ecosystems. Furthermore, cybersecurity professionals and stakeholders need to foster awareness and understanding of adversarial threats to mitigate risks and maintain the resilience of digital twin infrastructures.
- *Potential Framework Design:* The most crucial task for creating an AI-based cybersecurity system is to establish a solid foundation that supports automation, intelligence, and trustworthy decision-making. A well-designed AI framework for security modeling and experimental evaluation with DT data is both a very significant direction and a challenging problem. To do this, sophisticated algorithms need to be developed that can detect and mitigate cyber threats autonomously, taking advantage of machine learning techniques and natural language processing to analyze huge amounts of data generated by digital twins. A rigorous experimental evaluation, including diverse datasets, simulated attack scenarios, and real-world testing, is needed to ensure the reliability and effectiveness of such systems. To overcome these challenges, interdisciplinary collaboration and continuous innovation are required to advance the state-of-the-art in AI-driven cybersecurity, enhancing the resilience and security of digital infrastructures.

In summary, our study has revealed several potential future avenues for the study of cybersecurity in digital twins. First, more research needs to be done on the characteristics of DT data, including related features, data distributions, and pertinent contexts. Second, real-world evaluation of the scalability and effectiveness of current analytics methodologies applied to DT data is required. Thirdly, innovative methods and algorithms handling the underlying issues are needed to develop. Fourth, a range of empirical evaluations are necessary to quantify the performance of these AI techniques and to compare their efficacy and efficiency to those of currently used techniques. Fifth, additional effort is required to effectively deploy the ultimate models in a way that will achieve automation, intelligence, and trustworthiness in the relevant application domains. Overall, the concerns with research and prospective methods outlined above could help the community realize the full potential of AI/XAI-based cyber modeling in the digital twin environment. It will require continual research and development, as well as collaboration between cybersecurity professionals, AI specialists, and DT experts, to address the issues and capitalize on the potential provided by AI/XAI for next-generation cybersecurity in a digital twin environment.

7. Discussion

The above study and literature review assessed that methods based on AI/XAI have the potential to make significant advancements in a variety of application areas in the context of cybersecurity in the digital twin environment. The growing complexity of cybersecurity threats in DT and the increasing use of AI and data-driven technologies for security tasks have made it essential to develop transparent and interpretable models that can help analysts better understand the security landscape and make more informed decisions [9]. Leveraging machine learning algorithms and other AI technologies discussed in Section 4, offers several benefits, including proactive threat detection, improved resilience as well as interpretability analysis in the context of cybersecurity in the digital twin. Furthermore, XAI methods can support analysts and security professionals in comprehending how the system functions, identifying potential vulnerabilities, and ultimately leading to the development of trustworthy cyber systems with intelligent decision-making.

The integration of XAI into cybersecurity modeling in a digital twin environment can provide a better understanding of the behavior of the system and the potential threats it may face. One of the key advantages of using digital twin-based XAI models for cybersecurity is that they can provide a comprehensive view of the security landscape answering the questions - "How does DT enhance cybersecurity resilience using AI?" and "How AI can help to mitigate the possible threats and anomalies in DT", discussed in Section 3. By simulating different scenarios and analyzing the behavior of the system

under different conditions, it becomes possible to identify potential vulnerabilities and threats that might be missed by human analysts or other approaches. This can help analysts develop more effective security strategies and prioritize their resources more efficiently. While there have been significant advances in AI research in recent years [18], AI/XAI-based cyber modeling still faces many issues including available data and modeling algorithms, summarized and discussed in Section 6. Furthermore, making the AI-based cyber model explainable and obtaining the human-level accuracy of such systems is challenging and demanding for both researchers and practitioners. As XAI can build alternative systems, models, and algorithms with human-understandable capabilities, it is also challenging to decide which XAI methods could be useful to tackle a certain problem in DT. Typically, AI models depend on several factors, such as the nature of the problem, the available data, the computational resources, the interpretability requirements, and the specific project goals. Thus our AI/XAI taxonomy analysis discussed in Section 4 could be a potential source and guidelines to support the researchers and practitioners in this emerging area of study. There is still much work to be done to develop techniques that can provide reliable and understandable explanations of complex AI, ML, LLM or other black-box models for their effective and trustable use in real-world application areas.

Overall, AI and XAI have become major cyber industry concerns in DT. The future of almost all contexts and humanity including future social context, safety and security, and eventually the quality of human life will be impacted by these emerging technologies as they continue to grow. Thus, AI/XAI-based modeling can substantially progress and open up new horizons to advance the cyberspace in DT, and eventually can lead to the next-generation cybersecurity systems in a digital twin environment.

8. Conclusion

Motivated by the need for cybersecurity automation, intelligence, and trustworthiness, this article provided an extensive study and synthesis on AI/XAI-based modeling in DT. The study began with formulating key questions accordingly and how these methods can be employed to resolve various real-world cyber issues in digital twin environments. Next how it can enhance cyber resilience as well as security in different layers of digital twin are highlighted to go forward with AI-based modeling. We then provided a taxonomy with a thorough study of recent advances in AI and XAI-based methods with their potential usability in the context of cybersecurity. Several use cases such as predictive maintenance, intrusion detection, access control, cyber awareness generation, etc. were highlighted to inspire and provide a clear picture and understanding of the potentiality of AI/XAI-based cybersecurity modeling in DT. Hence, AI is the key element of cybersecurity enhancement that enables systems to carry out activities automatically and intelligently, whereas XAI provides a collection of processes that can produce explanations from different perspectives that are understandable by human analysts. Finally, several key challenges and prospects have been identified and discussed based on our study that may aid the community in understanding and realizing the full potential of AI-based cyber modeling in this emerging area of study. We believe our study and in-depth analysis from the perspective of automation, intelligence, and trustworthiness might serve as a reference guide and foundation for researchers and industry professionals, as well as policy makers, and provide a roadmap for the next-generation cybersecurity applications.

CRediT authorship contribution statement

Iqbal H. Sarker: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Visualization, Conceptualization. **Helge Janicke:** Writing – review & editing, Adviser. **Ahmad Mohsin:** Writing – review & editing. **Asif Gill:** Writing – review & editing. **Leandros Maglaras:** Writing – review & editing.

Declaration of competing interest

The authors declare no conflict of interests.

Acknowledgments

“The work has been supported by the Cyber Security Research Centre Limited whose activities are partially funded by the Australian Government’s Cooperative Research Centres Program”.

References

[1] B.R. Barricelli, E. Casiraghi, D. Fogli, A survey on digital twin: Definitions, characteristics, applications, and design implications, IEEE Access 7 (2019) 167653–167671.

[2] C. Alcaraz, J. Lopez, Digital twin: A comprehensive survey of security threats, IEEE Commun. Surv. Tutor. (2022).

[3] G. Mylonas, A. Kalogeras, G. Kalogeras, C. Anagnostopoulos, C. Alexakos, L. Muñoz, Digital twins from smart manufacturing to smart cities: A survey, IEEE Access 9 (2021) 143222–143249.

[4] S.P. Ramu, P. Boopalan, Q.-V. Pham, P.K.R. Maddikunta, T. Huynh-The, M. Alazab, T.T. Nguyen, T.R. Gadekallu, Federated learning enabled digital twins for smart cities: Concepts, recent advances, and future directions, Sustainable Cities Soc. 79 (2022) 103663.

[5] B. Sousa, M. Arieiro, V. Pereira, J. Correia, N. Lourenço, T. Cruz, ELEGANT: Security of critical infrastructures with digital twins, IEEE Access 9 (2021) 107574–107588.

[6] H. Elayan, M. Aloqaily, M. Guizani, Digital twin for intelligent context-aware IoT healthcare systems, IEEE Internet Things J. 8 (23) (2021) 16749–16757.

[7] W. Purcell, T. Neubauer, Digital twins in agriculture: A state-of-the-art review, Smart Agric. Technol. (2022) 100094.

[8] A.K. Sleiti, J.S. Kapat, L. Vesely, Digital twin in energy industry: Proposed robust digital twin for power plant and other complex capital-intensive large engineering systems, Energy Rep. 8 (2022) 3704–3726.

[9] I.H. Sarker, AI-Driven Cybersecurity and Threat Intelligence: Cyber Automation, Intelligent Decision-Making and Explainability, Springer, 2024.

[10] N. Kaloudi, J. Li, The ai-based cyber threat landscape: A survey, ACM Comput. Surv. 53 (1) (2020) 1–34.

[11] I.H. Sarker, H. Janicke, M.A. Ferrag, A. Abuadbba, Multi-aspect rule-based AI: Methods, taxonomy, challenges and directions toward automation, intelligence and transparent cybersecurity modeling for critical infrastructures, Internet Things (2024) 101110.
[12] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.
[13] R. Ibrahim, M.O. Shafiq, Explainable convolutional neural networks: A taxonomy, review, and future directions, ACM Comput. Surv. 55 (10) (2023) 1–37.
[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. (CSUR) 51 (5) (2018) 1–42.
[15] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Comput. Surv. 55 (9) (2023) 1–33.
[16] N. Capuano, G. Fenza, V. Loia, C. Stanzione, Explainable artificial intelligence in CyberSecurity: A survey, IEEE Access 10 (2022) 93575–93600.
[17] D. Wagg, K. Worden, R. Barthorpe, P. Gardner, Digital twins: state-of-the-art and future directions for modeling and simulation in engineering dynamics applications, ASCE-ASME J. Risk Uncertain. Engrg. Syst. B Mech. Engrg. 6 (3) (2020).
[18] I.H. Sarker, Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems, SN Comput. Sci. 3 (2) (2022) 158.
[19] M.M. Rathore, S.A. Shah, D. Shukla, E. Bentafat, S. Bakiras, The role of ai, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities, IEEE Access 9 (2021) 32030–32052.
[20] Y. Hu, W. Kuang, Z. Qin, K. Li, J. Zhang, Y. Gao, W. Li, K. Li, Artificial intelligence security: Threats and countermeasures, ACM Comput. Surv. 55 (1) (2021) 1–36.
[21] D. Kaur, S. Uslu, K.J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM Comput. Surv. 55 (2) (2022) 1–38.
[22] M. Kuzlu, C. Fair, O. Guler, Role of artificial intelligence in the Internet of Things (IoT) cybersecurity, Discov. Internet Things 1 (2021) 1–14.
[23] S. Samtani, M. Kantarcioglu, H. Chen, Trailblazing the artificial intelligence for cybersecurity discipline: a multi-disciplinary research roadmap, ACM Trans. Manag. Inf. Syst. (TMIS) 11 (4) (2020) 1–19.
[24] M. Alazab, S.P. RM, M. Parimala, P.K.R. Maddikunta, T.R. Gadekallu, Q.-V. Pham, Federated learning for cybersecurity: concepts, challenges, and future directions, IEEE Trans. Ind. Inform. 18 (5) (2021) 3501–3509.
[25] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[26] M. SEALE, Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities, IEEE Access (2022).
[27] A. Rawal, J. McCoy, D.B. Rawat, B.M. Sadler, R.S. Amant, Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives, IEEE Trans. Artif. Intell. 3 (6) (2021) 852–866.
[28] F. Charmet, H.C. Tanuwidjaja, S. Ayoubi, P.-F. Gimenez, Y. Han, H. Jmila, G. Blanc, T. Takahashi, Z. Zhang, Explainable artificial intelligence for cybersecurity: a literature survey, Ann. Telecommun. (2022) 1–24.
[29] I. Ahmed, G. Jeon, F. Piccialli, From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where, IEEE Trans. Ind. Inform. 18 (8) (2022) 5031–5042.
[30] W. Saeed, C. Omlin, Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities, Knowl.-Based Syst. (2023) 110273.
[31] I.H. Sarker, Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview, Secur. Privacy (2022) e295.
[32] E. Bout, V. Loscri, A. Gallais, How machine learning changes the nature of cyberattacks on IoT networks: A survey, IEEE Commun. Surv. Tutor. 24 (1) (2021) 248–279.
[33] R. Faleiro, L. Pan, S.R. Pokhrel, R. Doss, Digital twin for cybersecurity: Towards enhancing cyber resilience, in: Broadband Communications, Networks, and Systems: 12th EAI International Conference, BROADNETS 2021, Virtual Event, October 28–29, 2021, Proceedings 12, Springer, 2022, pp. 57–76.
[34] D. Holmes, M. Papathanasaki, L. Maglaras, M.A. Ferrag, S. Nepal, H. Janicke, Digital twins and cyber security–solution or challenge? in: 2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM, IEEE, 2021, pp. 1–8.
[35] I.H. Sarker, Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions, SN Comput. Sci. 2 (6) (2021) 420.
[36] I.H. Sarker, CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks, Internet Things 14 (2021) 100393.
[37] G. Apruzzese, P. Laskov, E. Montes de Oca, W. Mallouli, L. Brdalo Rapa, A.V. Grammatopoulos, F. Di Franco, The role of machine learning in cybersecurity, Digit. Threat.: Res. Pract. 4 (1) (2023) 1–38.
[38] F. Tao, H. Zhang, A. Liu, A.Y. Nee, Digital twin in industry: State-of-the-art, IEEE Trans. Ind. Inform. 15 (4) (2018) 2405–2415.
[39] Z. Zhang, H. Al Hamadi, E. Damiani, C.Y. Yeun, F. Taher, Explainable artificial intelligence applications in cyber security: State-of-the-art in research, IEEE Access (2022).
[40] A. Humayed, J. Lin, F. Li, B. Luo, Cyber-physical systems security—A survey, IEEE Internet Things J. 4 (6) (2017) 1802–1831.
[41] I.H. Sarker, A.I. Khan, Y.B. Abushark, F. Alsolami, Internet of Things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions, Mob. Netw. Appl. (2022) 1–17.
[42] S. Kim, K.-J. Park, C. Lu, A survey on network security for cyber–physical systems: From threats to resilient design, IEEE Commun. Surv. Tutor. 24 (3) (2022) 1534–1573.
[43] I.H. Sarker, Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects, Ann. Data Sci. (2022) 1–26.
[44] A. Halbouni, T.S. Gunawan, M.H. Habaebi, M. Halbouni, M. Kartiwi, R. Ahmad, Machine learning and deep learning approaches for cybersecuriy: A review, IEEE Access (2022).
[45] I. Rosenberg, A. Shabtai, Y. Elovici, L. Rokach, Adversarial machine learning attacks and defense methods in the cyber security domain, ACM Comput. Surv. 54 (5) (2021) 1–36.
[46] K. He, D.D. Kim, M.R. Asghar, Adversarial machine learning for network intrusion detection systems: A comprehensive survey, IEEE Commun. Surv. Tutor. (2023).
[47] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, Q. Yu, A survey of adversarial attack and defense methods for malware classification in cyber security, IEEE Commun. Surv. Tutor. (2022).
[48] E. Zhu, Y. Ju, Z. Chen, F. Liu, X. Fang, DTOF-ANN: an artificial neural network phishing detection model based on decision tree and optimal features, Appl. Soft Comput. 95 (2020) 106505.
[49] Y. Chai, Y. Zhou, W. Li, Y. Jiang, An explainable multi-modal hierarchical attention model for developing phishing threat intelligence, IEEE Trans. Dependable Secure Comput. 19 (2) (2021) 790–803.
[50] I.H. Sarker, Y.B. Abushark, F. Alsolami, A.I. Khan, Intrudtree: a machine learning based cyber security intrusion detection model, Symmetry 12 (5) (2020) 754.

[51] M. Wang, K. Zheng, Y. Yang, X. Wang, An explainable machine learning framework for intrusion detection systems, IEEE Access 8 (2020) 73127–73141.
[52] I. Psychoula, A. Gutmann, P. Mainali, S.H. Lee, P. Dunphy, F. Petitcolas, Explainable machine learning for fraud detection, Computer 54 (10) (2021) 49–59.
[53] A. Barbado, Ó. Corcho, R. Benjamins, Rule extraction in unsupervised anomaly detection for model explainability: Application to OneClass SVM, Expert Syst. Appl. 189 (2022) 116100.
[54] T. Dias, N. Oliveira, N. Sousa, I. Praça, O. Sousa, A hybrid approach for an interpretable and explainable intrusion detection system, in: Intelligent Systems Design and Applications: 21st International Conference on Intelligent Systems Design and Applications, ISDA 2021 Held During December 13–15, 2021, Springer, 2022, pp. 1035–1045.
[55] Z. Pan, J. Sheldon, P. Mishra, Hardware-assisted malware detection and localization using explainable machine learning, IEEE Trans. Comput. 71 (12) (2022) 3308–3321.
[56] Y. Lin, R. Liu, D.M. Divakaran, J.Y. Ng, Q.Z. Chan, Y. Lu, Y. Si, F. Zhang, J.S. Dong, Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages, in: USENIX Security Symposium, 2021, pp. 3793–3810.
[57] B. Wu, S. Chen, C. Gao, L. Fan, Y. Liu, W. Wen, M.R. Lyu, Why an android app is classified as malware: Toward malware classification interpretation, ACM Trans. Softw. Eng. Methodol. (TOSEM) 30 (2) (2021) 1–29.
[58] G. Iadarola, F. Martinelli, F. Mercaldo, A. Santone, Towards an interpretable deep learning model for mobile malware detection and family identification, Comput. Secur. 105 (2021) 102198.
[59] F. Ullah, H. Naeem, S. Jabbar, S. Khalid, M.A. Latif, F. Al-Turjman, L. Mostarda, Cyber security threats detection in internet of things using deep learning approach, IEEE Access 7 (2019) 124379–124389.
[60] C. Joshi, R.K. Ranjan, V. Bharti, A fuzzy logic based feature engineering approach for botnet detection using ANN, J. King Saud Univ.-Comput. Inf. Sci. 34 (9) (2022) 6872–6882.
[61] P.P. Kundu, T. Truong-Huu, L. Chen, L. Zhou, S.G. Teo, Detection and classification of botnet traffic using deep learning with model explanation, IEEE Trans. Dependable Secure Comput. (2022).
[62] I. Dimitriadis, K. Georgiou, A. Vakali, Social botomics: A systematic ensemble ml approach for explainable and multi-class bot detection, Appl. Sci. 11 (21) (2021) 9857.
[63] C.S. Wickramasinghe, K. Amarasinghe, D.L. Marino, C. Rieger, M. Manic, Explainable unsupervised machine learning for cyber-physical systems, IEEE Access 9 (2021) 131824–131843.
[64] K. Shaukat, S. Luo, V. Varadharajan, I.A. Hameed, M. Xu, A survey on machine learning techniques for cyber security in the last decade, IEEE Access 8 (2020) 222310–222354.
[65] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, C. Wang, Machine learning and deep learning methods for cybersecurity, IEEE Access 6 (2018) 35365–35381.
[66] I.H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Comput. Sci. 2 (3) (2021) 160.
[67] M. Cui, J. Wang, M. Yue, Machine learning-based anomaly detection for load forecasting under cyberattacks, IEEE Trans. Smart Grid 10 (5) (2019) 5724–5734.
[68] M.A. Bouke, A. Abdullah, S.H. ALshatebi, M.T. Abdullah, E2IDS: An enhanced intelligent intrusion detection system based on decision tree algorithm, J. Appl. Artif. Intell. 3 (1) (2022) 1–16.
[69] R. Heartfield, G. Loukas, A. Bezemskij, E. Panaousis, Self-configurable cyber-physical intrusion detection for smart homes using reinforcement learning, IEEE Trans. Inf. Forensics Secur. 16 (2020) 1720–1735.
[70] J. Liang, Z. Qin, S. Xiao, L. Ou, X. Lin, Efficient and secure decision tree classification for cloud-assisted online diagnosis services, IEEE Trans. Dependable Secure Comput. 18 (4) (2019) 1632–1644.
[71] M.R.C. Acosta, S. Ahmed, C.E. Garcia, I. Koo, Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks, IEEE Access 8 (2020) 19921–19933.
[72] Z. Lv, Y. Li, H. Feng, H. Lv, Deep learning for security in digital twins of cooperative intelligent transportation systems, IEEE Trans. Intell. Transp. Syst. 23 (9) (2021) 16666–16675.
[73] Y. Luo, Y. Xiao, L. Cheng, G. Peng, D. Yao, Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities, ACM Comput. Surv. 54 (5) (2021) 1–36.
[74] W. Danilczyk, Y.L. Sun, H. He, Smart grid anomaly detection using a deep learning digital twin, in: 2020 52nd North American Power Symposium, NAPS, IEEE, 2021, pp. 1–6.
[75] H. HaddadPajouh, A. Dehghantanha, R. Khayami, K.-K.R. Choo, A deep recurrent neural network based approach for internet of things malware threat hunting, Future Gener. Comput. Syst. 85 (2018) 88–96.
[76] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Vol. 1215, 1994, pp. 487–499.
[77] J.R. Quinlan, C4.5: Programs for machine learning, Mach. Learn. (1993).
[78] M. Hasanipanah, H. Bakhshandeh Amnieh, A fuzzy rule-based approach to address uncertainty in risk assessment and prediction of blast-induced flyrock in a quarry, Nat. Resour. Res. 29 (2020) 669–689.
[79] M. Alali, A. Almogren, M.M. Hassan, I.A. Rassan, M.Z.A. Bhuiyan, Improving risk assessment model of cyber security using fuzzy logic inference system, Comput. Secur. 74 (2018) 323–339.
[80] Z.-J. Zhou, G.-Y. Hu, C.-H. Hu, C.-L. Wen, L.-L. Chang, A survey of belief rule-base expert system, IEEE Trans. Syst. Man Cybern.: Syst. 51 (8) (2019) 4944–4958.
[81] R. Ul Islam, M.S. Hossain, K. Andersson, A novel anomaly detection algorithm for sensor data under uncertainty, Soft Comput. 22 (5) (2018) 1623–1639.
[82] I.H. Sarker, LLM potentiality and awareness: A position paper from the perspective of trustworthy and responsible AI modeling, 2024, Authorea Preprints.
[83] R. Sharma, R. Sibal, S. Sabharwal, Software vulnerability prioritization using vulnerability description, Int. J. Syst. Assur. Eng. Manag. 12 (2021) 58–64.
[84] J.S. Garrido, D. Dold, J. Frank, Machine learning on knowledge graphs for context-aware security monitoring, in: 2021 IEEE International Conference on Cyber Security and Resilience, CSR, IEEE, 2021, pp. 55–60.
[85] A. Piplai, S. Mittal, A. Joshi, T. Finin, J. Holt, R. Zak, Creating cybersecurity knowledge graphs from malware after action reports, IEEE Access 8 (2020) 211691–211703.
[86] S. Wang, J. Wan, D. Li, C. Liu, Knowledge reasoning with semantic data for real-time data processing in smart factory, Sensors 18 (2) (2018) 471.
[87] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.
[88] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (Suppl 1) (2023) 1513–1589.
[89] J. Zhang, J. Yin, R. Wang, Basic framework and main methods of uncertainty quantification, Math. Probl. Eng. 2020 (2020) 1–18.
[90] M. Liggins II, D. Hall, J. Llinas, Handbook of Multisensor Data Fusion: Theory and Practice, CRC Press, 2017.
[91] Z.H. Qaisar, R. Li, Multimodal information fusion for android malware detection using lazy learning, Multimedia Tools Appl. (2022) 1–15.
[92] S. Dey, Q. Ye, S. Sampalli, A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks, Inf. Fusion 49 (2019) 205–215.

[93] M.J. Kaur, V.P. Mishra, P. Maheshwari, The convergence of digital twin, IoT, and machine learning: transforming data into action, Digit. Twin Technol. Smart Cities (2020) 3–17.

[94] S. Hariharan, R. Rejimol Robinson, R.R. Prasad, C. Thomas, N. Balakrishnan, XAI for intrusion detection system: comparing explanations based on global and local scope, J. Comput. Virol. Hack. Tech. (2022) 1–23.

[95] D. Szafron, B. Poulin, R. Eisner, P. Lu, R. Greiner, D. Wishart, A. Fyshe, B. Pearcy, C. Macdonell, J. Anvik, Visual explanation of evidence in additive classifiers, in: Proceedings of Innovative Applications of Artificial Intelligence, Vol. 2, 2006.

[96] H. Chen, X. Chen, S. Shi, Y. Zhang, Generate natural language explanations for recommendation, 2021, arXiv preprint arXiv:2101.03392.

[97] H. Liu, Q. Yin, W.Y. Wang, Towards explainable NLP: A generative explanation framework for text classification, 2018, arXiv preprint arXiv:1811.00196.

[98] I. Sarker, A. Colman, J. Han, P. Watters, Context-Aware Machine Learning and Mobile Data Analytics: Automated Rule-Based Services with Intelligent Decision-Making, Springer, 2021.

[99] R. Langone, A. Cuzzocrea, N. Skantzos, Interpretable anomaly prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools, Data Knowl. Eng. 130 (2020) 101850.

[100] N. Mehdiyev, P. Fettke, Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring, Interpret. Artif. Intell.: A Perspect. Granul. Comput. (2021) 1–28.

[101] S.S.S. Sindhu, S. Geetha, A. Kannan, Decision tree based light weight intrusion detection using a wrapper approach, Expert Syst. Appl. 39 (1) (2012) 129–141.

[102] I.H. Sarker, Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective, SN Comput. Sci. 2 (5) (2021) 377.

[103] G. Baryannis, S. Dani, G. Antoniou, Predicting supply chain risks using machine learning: The trade-off between performance and interpretability, Future Gener. Comput. Syst. 101 (2019) 993–1004.

[104] A. Okutan, S.J. Yang, K. McConky, Predicting cyber attacks with bayesian networks using unconventional signals, in: Proceedings of the 12th Annual Conference on Cyber and Information Security Research, 2017, pp. 1–4.

[105] X. Fang, M. Xu, S. Xu, P. Zhao, A deep learning framework for predicting cyber attacks rates, EURASIP J. Inf. Secur. 2019 (2019) 1–11.

[106] A. Castellani, S. Schmitt, S. Squartini, Real-world anomaly detection by using digital twin systems and weakly supervised learning, IEEE Trans. Ind. Inform. 17 (7) (2020) 4733–4742.

[107] E.C. Balta, M. Pease, J. Moyne, K. Barton, D.M. Tilbury, Digital twin-based cyber-attack detection framework for cyber-physical manufacturing systems, IEEE Trans. Autom. Sci. Eng. (2023).

[108] Q. Xu, S. Ali, T. Yue, Digital twin-based anomaly detection in cyber-physical systems, in: 2021 14th IEEE Conference on Software Testing, Verification and Validation, ICST, IEEE, 2021, pp. 205–216.

[109] O.K. Sahingoz, E. Buber, O. Demir, B. Diri, Machine learning based phishing detection from URLs, Expert Syst. Appl. 117 (2019) 345–357.

[110] J. Qiu, J. Zhang, W. Luo, L. Pan, S. Nepal, Y. Xiang, A survey of android malware detection with deep neural models, ACM Comput. Surv. 53 (6) (2020) 1–36.

[111] G. Kocher, G. Kumar, Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges, Soft Comput. 25 (15) (2021) 9731–9763.

[112] M. Shafiq, Z. Tian, A.K. Bashir, X. Du, M. Guizani, Corrauc: a malicious bot-IoT traffic detection method in IoT network using machine-learning techniques, IEEE Internet Things J. 8 (5) (2020) 3242–3254.

[113] G. Vallathan, A. John, C. Thirumalai, S. Mohan, G. Srivastava, J.C.-W. Lin, Suspicious activity detection using deep learning in secure assisted living IoT environments, J. Supercomput. 77 (2021) 3242–3260.

[114] S. Garg, K. Kaur, N. Kumar, J.J. Rodrigues, Hybrid deep-learning-based anomaly detection scheme for suspicious flow detection in SDN: A social multimedia perspective, IEEE Trans. Multimed. 21 (3) (2019) 566–578.

[115] K. Vidović, I. Tomičić, K. Slovenec, M. Mikuc, I. Brajdić, Ranking network devices for alarm prioritisation: Intrusion detection case study, in: 2021 International Conference on Software, Telecommunications and Computer Networks, SoftCOM, IEEE, 2021, pp. 1–5.

[116] Q. Yan, M. Wang, W. Huang, X. Luo, F.R. Yu, Automatically synthesizing DoS attack traces using generative adversarial networks, Int. J. Mach. Learn. Cybern. 10 (12) (2019) 3387–3396.

[117] S. Mouti, S.K. Shukla, S. Althubiti, M.A. Ahmed, F. Alenezi, M. Arumugam, Cyber security risk management with attack detection frameworks using multi connect variational auto-encoder with probabilistic Bayesian networks, Comput. Electr. Eng. 103 (2022) 108308.

[118] M. Ibrahim, R. Elhafiz, Modeling an intrusion detection using recurrent neural networks, J. Eng. Res. 11 (1) (2023) 100013.

[119] J. Yin, M. Tang, J. Cao, H. Wang, Apply transfer learning to cybersecurity: Predicting exploitability of vulnerabilities by description, Knowl.-Based Syst. 210 (2020) 106529.

[120] I. Zografopoulos, J. Ospina, X. Liu, C. Konstantinou, Cyber-physical energy systems security: Threat modeling, risk assessment, resources, metrics, and case studies, IEEE Access 9 (2021) 29775–29818.

[121] M. Dietz, M. Vielberth, G. Pernul, Integrating digital twin security simulations in the security operations center, in: Proceedings of the 15th International Conference on Availability, Reliability and Security, 2020, pp. 1–9.

[122] C. Gehrmann, M. Gunnarsson, A digital twin based industrial automation and control system security architecture, IEEE Trans. Ind. Inform. 16 (1) (2019) 669–680.

[123] J. Heaps, R. Krishnan, Y. Huang, J. Niu, R. Sandhu, Access control policy generation from user stories using machine learning, in: Data and Applications Security and Privacy XXXV: 35th Annual IFIP WG 11.3 Conference, DBSec 2021, Calgary, Canada, July 19–20, 2021, Proceedings 35, Springer, 2021, pp. 171–188.

[124] M.N. Nobi, M. Gupta, L. Praharaj, M. Abdelsalam, R. Krishnan, R. Sandhu, Machine learning in access control: A taxonomy and survey, 2022, arXiv preprint arXiv:2207.01739.

[125] Y. Liu, M. Dong, K. Ota, J. Li, J. Wu, Deep reinforcement learning based smart mitigation of ddos flooding in software-defined networks, in: 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD, IEEE, 2018, pp. 1–6.

[126] H. Alturkistani, M.A. El-Affendi, Optimizing cybersecurity incident response decisions using deep reinforcement learning, Int. J. Electr. Comput. Eng. 12 (6) (2022) 6768.

[127] K. Hughes, K. McLaughlin, S. Sezer, Policy-based profiles for network intrusion response systems, in: 2022 IEEE International Conference on Cyber Security and Resilience, CSR, IEEE, 2022, pp. 279–286.

[128] M. Bashendy, A. Tantawy, A. Erradi, Intrusion response systems for cyber-physical systems: A comprehensive survey, Comput. Secur. (2022) 102984.

[129] B. Steenwinckel, D. De Paepe, S.V. Hautte, P. Heyvaert, M. Bentefrit, P. Moens, A. Dimou, B. Van Den Bossche, F. De Turck, S. Van Hoecke, et al., FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing

expert knowledge with machine learning, Future Gener. Comput. Syst. 116 (2021) 30–48.
[130] M. Eckhart, A. Ekelhart, E. Weippl, Enhancing cyber situational awareness for cyber-physical systems through digital twins, in: 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, IEEE, 2019, pp. 1222–1225.
[131] M. Groshev, C. Guimarães, J. Martín-Pérez, A. de la Oliva, Toward intelligent cyber-physical systems: Digital twin meets artificial intelligence, IEEE Commun. Mag. 59 (8) (2021) 14–20.
[132] M.A. Ferrag, O. Friha, L. Maglaras, H. Janicke, L. Shu, Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis, IEEE Access 9 (2021) 138509–138542.
[133] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B.B. Gupta, X. Chen, X. Wang, A survey of deep active learning, ACM Comput. Surv. (CSUR) 54 (9) (2021) 1–40.
[134] J.Z. Bengar, J. van de Weijer, B. Twardowski, B. Raducanu, Reducing label effort: Self-supervised meets active learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1631–1639.
[135] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE Trans. Knowl. Data Eng. 35 (1) (2021) 857–876.
[136] M. Mohammadpourfard, Y. Weng, M. Pechenizkiy, M. Tajdinian, B. Mohammadi-Ivatloo, Ensuring cybersecurity of smart grid against data integrity attacks under concept drift, Int. J. Electr. Power Energy Syst. 119 (2020) 105947.
[137] B. Bayram, B. Köroğlu, M. Gönen, Improving fraud detection and concept drift adaptation in credit card transactions using incremental gradient boosting trees, in: 2020 19th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2020, pp. 545–550.
[138] Z. Li, W. Huang, Y. Xiong, S. Ren, T. Zhu, Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm, Knowl.-Based Syst. 195 (2020) 105694.
[139] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under concept drift: A review, IEEE Trans. Knowl. Data Eng. 31 (12) (2018) 2346–2363.
[140] I.H. Sarker, A. Colman, J. Han, Recencyminer: mining recency-based personalized behavior from contextual smartphone data, J. Big Data 6 (1) (2019) 1–21.
[141] G. Siewruk, W. Mazurczyk, Context-aware software vulnerability classification using machine learning, IEEE Access 9 (2021) 88852–88867.
[142] L.F. Sikos, Cybersecurity knowledge graphs, Knowl. Inf. Syst. (2023) 1–21.

Iqbal H. Sarker received his Ph.D. in Computer Science from Swinburne University of Technology, Melbourne, Australia in 2018. Now he is working as a Research Fellow of the Cyber Security Cooperative Research Centre (CRC) in association with the Centre for Securing Digital Futures, Edith Cowan University (ECU), Australia. His research interests include Cybersecurity, AI/XAI and Machine Learning Algorithms, Data Science and Behavioral Analytics, Trustworthy LLMs, Knowledge and Rule Mining, Digital Twin, Critical Infrastructures and Industrial Applications. He has published 100+ journal and conference papers in various reputed venues published by Elsevier, Springer Nature, IEEE, ACM, Oxford University Press, etc. Moreover, he is a lead author of the books "Context-Aware Machine Learning and Mobile Data Analytics", and "AI-driven Cybersecurity and Threat Intelligence", published by Springer Nature, Switzerland. He has also been listed in the world's top 2% of most-cited scientists, published by Elsevier & Stanford University, USA. In addition to research work and publications, Dr. Sarker is also involved in a number of research engagement and leadership roles such as Journal editorial, international conference program committee (PC), student supervision, visiting scholar and national/international collaboration. He is a member of IEEE, ACM and Australian Information Security Association.

Helge Janicke is a Professor of Cybersecurity at Edith Cowan University (ECU), Australia. He is the Director of ECU's Security Research Institute and the Research Director for Australia's Cyber Security Cooperative Research Centre. He optained his PhD in 2007 from De Montfort University, UK, where he established DMU's Cyber Technology Institute and its Airbus Centre of Excellence for SCADA cybersecurity and digital forensics research, as well as heading up DMU's School of Computer Science. His research interests are Cybersecurity in Critical Infrastructure, Human Factors of Cybersecurity, Cybersecurity of Emerging Technologies, Digital Twins and Industrial IoT.

Ahmad Mohsin is a Post-doctoral Research Fellow of the Cyber Security Cooperative Research Centre (CSCRC) in association with the ECU Security Research Institute, Edith Cowan University (ECU), Australia. He completed his Ph.D. in Computers Science and Software Engineering from ECU, Australia. He has worked closely with Academia and Industry for designing intelligent and trustworthy systems. Ahmad has actively contributed to research projects in the Australian industry specific to Critical Infrastructures. His research interests include Cybersecurity, Predictive modeling of smart systems, Trustworthy and Reliable AI, Digital Twins and Critical Infrastructure Resilience. Ahmad has publications both in top Journals and conferences and is also an active reviewer of leading journals and conferences. Ahmad is a regular guest speaker at various academic and industry avenues.

Asif Gill is A/Professor and Head of Discipline Software Engineering at the School of Computer Science, UTS. He is also a Director of the DigiSAS Lab. He has PhD Computing, MSc Computing and Master of Business. His work focuses on architecting, implementing and evaluating the information-driven large scale secure and sustainable digital ecosystems. He is a member of the ACS Data Sharing Committee, IFIP Technical Committee 8.1, and Standards Australia Software and Systems Engineering Committee IT-015. He is often invited and involved as a professional keynote speaker, editor, conference chair, organizer and reviewer for a number of national and international academic and industry conferences.

Leandros A. Maglaras is a professor of cybersecurity in the School of Computing at Edinburgh Napier University. From September 2017 to November 2019, he was the Director of the National Cyber Security Authority of Greece. He obtained a B.Sc. (M.Sc. equivalent) in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 1998, M.Sc. in Industrial Production and Management from the University of Thessaly in 2004, and M.Sc. and Ph.D. degrees in Electrical & Computer Engineering from the University of Thessaly, in 2008 and 2014 respectively. In 2018 he was awarded a Ph.D. in Intrusion Detection in SCADA systems from the University of Huddersfield He is featured in Stanford University's list of the world's Top 2% scientists. He is a Senior Member of the Institute of Electrical & Electronics Engineers (IEEE) and is an author of more than 200 papers in scientific magazines and conferences.