 doi:10.20944/preprints202101.0457.v1

AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions

Iqbal H. Sarker[1,2*]

Abstract Artificial Intelligence (AI) is one of the key technologies of the Fourth Industrial Revolution (Industry 4.0), which can be used for the protection of Internet-connected systems from cyber-threats, attacks, damage, or unauthorized access. To intelligently solve today's various cybersecurity issues, popular AI techniques involving Machine Learning (ML) and Deep Learning (DL) methods, the concept of Natural Language Processing (NLP), Knowledge Representation and Reasoning (KRR), as well as the concept of knowledge or rule-based Expert Systems (ES) modeling can be used. Based on these AI methods, in this paper, we present a comprehensive view on *"AI-driven Cybersecurity"* that can play an important role for intelligent cybersecurity services and management. The security intelligence modeling based on such AI methods can make the cybersecurity computing process automated and intelligent than the conventional security systems. We also highlight several *research directions* within the scope of our study, which can help researchers do future research in the area. Overall, this paper's ultimate objective is to serve as a reference point and guidelines for cybersecurity researchers as well as industry professionals in the area, especially from an AI-based technical point of view.

Keywords Cybersecurity; artificial intelligence; machine learning; cyber data analytics; cyber-attacks; anomaly; intrusion detection; security intelligence.

[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia.
[2] Chittagong University of Engineering and Technology, Chittagong 4349, Bangladesh.

*Corresponding: msarker@swin.edu.au (Iqbal H. Sarker)

1 Introduction

The modern world depends more on technology than ever before. A huge amount of data is generated and gathered with the large implementation of booming technologies such as the Internet of Things (IoT) [95] and cloud computing [162]. Although data can be used to better serve the corresponding business needs, cyber-attacks often pose major challenges. A cyber-attack is usually a malicious and concerted attempt by an individual or organization to breach another individual or organization's information system. Malware attack, ransomware, denial of service (DoS), phishing or social engineering, SQL injection attack, Man-in-the-middle, Zero-day exploit, or insider threats are common nowadays in the area [141]. These types of security incidents or cybercrime can affect organizations and individuals, cause disruptions, as well as devastating financial losses. For instance, according to the IBM report, a data breach costs 8.19 million USD for the United States [15], and the estimated annual cost to the global economy from cybercrime is 400 billion USD [60]. Cybercrimes are growing at an exponential rate that brings an alarming message for the cybersecurity professionals and researchers [141]. Therefore, to effectively and intelligently protect an information system, particularly, Internet-connected systems from various cyber-threats, attacks, damage, or unauthorized access, is a key issue to be solved urgently, in which we are interested in this paper.

In the real world, the overall national security of the business, government, organizations, and individual citizens of a country depends on the security management tools having the capability of detecting and preventing the security incidents in a timely and intelligent way. Intelligent cybersecurity services and management are

therefore essential because immense amounts of data on computers and other devices are collected, processed, and stored by government, military, corporate, financial, medical organizations, and many others. Cybersecurity usually refers to a collection of technologies, procedures, and practices designed to protect networks, computers, programs, and data from attack, disruption, or unauthorized access. It's also known as "information technology security" or "electronic information security". Several related terms with the concept of cybersecurity are briefly discussed and summarized in Section 2. According to today's numerous needs, the conventional well-known security solutions such as antivirus, firewalls, user authentication, encryption etc. may not be effective [36] [112] [157] [159]. The key problem with these traditional systems is that they are normally operated by a few experienced security experts, where data processing is carried out in an ad-hoc manner and can therefore not run intelligently according to needs [61] [143]. On the other hand, Artificial Intelligence (AI), which is known as the key technologies of the Fourth Industrial Revolution (Industry 4.0), can play an important role for intelligent cybersecurity services and management according to its computing power and capabilities. Thus, we focus on *"AI-driven Cybersecurity"* to make the cybersecurity computing process automated and intelligent than the conventional security systems in the area.

Artificial Intelligence (AI) is the branch of computer sciences that usually emphasizes the creation of intelligent machines, thinking and functioning like humans. To intelligently solve today's various cybersecurity issues, e.g., intrusion detection and prevention system, popular AI techniques involving Machine Learning (ML) and Deep Learning (DL) methods, the concept of Natural Language Processing (NLP), Knowledge Representation and Reasoning (KRR), as well as the concept of knowledge or rule-based Expert Systems (ES) modeling can be used, which are briefly discussed in Section 3. For instance, these techniques can be applied for identifying malicious activities, fraud detection, predicting cyber-attacks, access control management, detecting cyber-anomalies or intrusions, etc. The aim of this paper is therefore to provide a reference guide for those processionals from *academia and industry* who want to work and research based on AI methods in the field of cybersecurity. Therefore, in the sense of cybersecurity, great emphasis is put on common AI-based methods and their applicability for solving today's diverse security issues. Overall, this paper provides a detailed view of *AI-driven cybersecurity* in terms of principles and modeling for intelligent and automated cybersecurity services and management through intelligent decision making by taking into account the benefits of AI methods.

The main contributions of this paper are therefore listed as follows:

- To provide a brief overview on the concept of *AI-driven cybersecurity* for intelligent cybersecurity services and management according to today's needs. For this, we first briefly review the related methods and systems in the context of cybersecurity to motivate our study as well as to make a position for the term AI-driven cybersecurity.
- To present *security intelligence modeling* where various AI-based methods such as machine and deep learning, natural language processing, knowledge representation and reasoning, as well as the knowledge or rule-based expert systems modeling are taken into account according to our goal.
- Finally, we discuss and highlight several *research directions* within the scope of our study, which can help the cybersecurity researchers to do future research in the area.

The rest of the paper is organized as follows. Section 2 provides a background and reviews the related work in this domain. In Section 3, we discuss how various AI techniques can be used for security intelligence modeling. In section 4, we discover and summarize several research issues and potential future directions, and finally, Section 5 concludes this paper.

2 Background and Related Work

In this section, we provide an overview of the relevant AI-driven cybersecurity technologies, including different types of cybersecurity incidents within the scope of our study.

2.1 Basic Security Properties and CIA Triad

Confidentiality, integrity, and availability, also known as the CIA triad, is a model usually designed to guide information security policies within an organization. Thus to understand the security policy, the CIA triad with the mentioned properties is important that are discussed as below.

- *Confidentiality* is a property of security policy that typically refers to protecting the information and systems from unauthorized parties. Confidentiality threat can typically target databases, application servers, and system administrators, and can be considered as "data theft".

- *Integrity* is another property of security policy that typically refers to prevent any kind of destruction or modification of information by unauthorized parties. Integrity threat typically includes finance related threat like altering financial data, stealing money, reroute deposit, or hijacking, and to damage of the organization trustworthiness, and can be considered as "data alteration".
- *Availability* is also considered as another property of security policy that typically refers to ensure the access of information systems or assets to an authorized party or entity in a reliable and timely manner. Availability threat typically includes denial of service, or physical destruction, and can be considered as "denial access of the data".

Overall, based on the CIA triad for the security policy discussed above, we can simply conclude that "Confidentiality" is limiting the data access, "Integrity" is ensuring the data is accurate, and "Availability" is making sure the accessibility of the data to the right entity.

2.2 Cybersecurity and Related Terms

Over the last half-century, our modern and digital society is highly integrated with information and communication technology (ICT). As the smart computing devices used in our daily life activities are mostly driven by global Internet connectivity, the associated risk of data breaches or cyber-attacks is increasing day by day. Thus, preventing and protecting the ICT systems from various kinds of advanced cyber-attacks or threats, is known as ICT security, becomes the major concern for our security professionals or policymakers in recent days [128]. ICT security refers to relevant incidents as well as measures, controls, and procedures applied by enterprises to ensure integrity, confidentiality, and availability of their data and systems. Cybersecurity is simply about securing things that are vulnerable through ICT. Although the term "Cybersecurity" is popular nowadays, several relevant terms such as "Information security", "Data security", "Network security", "Internet/IoT security" often get interchangeable and may create confusion among the readers as well as the professionals in the area. In the following, we define these terms and highlight their world-wide popularity score as well.

- *Data security* is all about securing data, which could be specific to data, typically in storage. Thus, data security can be defined as the prevention of unauthorized access, use, disruption, modification, or destruction of data in storage.
- *Information security* is the prevention of unauthorized access, use, disruption, modification, or destruction of information. Information security, in a sense, can be considered as a specific discipline under the cybersecurity umbrella that is the broader practice of defending IT assets from attacks or threats.
- *Network security* is usually the practice of preventing and tracking unauthorized access, misuse, alteration, or denial of services available to a computer network. It thus can be considered as a subset of cybersecurity, which typically protects the data flowing over the network.
- *Internet security* is a specific aspect of broader concepts such as cybersecurity and computer security, focusing on the specific risks and vulnerabilities of internet access and use. *IoT security* is another relevant term, is typically concerned with protecting Internet-enabled devices, i.e., Internet of Things (IoT) devices, that connect on wireless networks [28].

The above-mentioned security terms are related to *"Cybersecurity"*, which is the practice of defending computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks, cyber-threats, damage, or unauthorized access. Among these terms, the worldwide popularity of "cybersecurity" is higher than others and increasing day-by-day, which is shown in Fig. 1. The popularity trend in Fig. 1 is shown based on the data collected from Google Trends over the last five years [22]. According to Fig. 1, the popularity indication values for cybersecurity was low in 2016 and is increasing day-by-day. Thus, in this paper, we focus on the popular term "cybersecurity", which is the key to achieving the Fourth Industrial Revolution (Industry 4.0).

Many researchers defined cybersecurity in various ways. For instance, the diverse activities or policies that are taken into account to protect the ICT systems from threats or attacks is known as cybersecurity [60]. Craigen et al. defined "cybersecurity as a set of tools, practices, and guidelines that can be used to protect computer networks, software programs, and data from attack, damage, or unauthorized access" [51]. According to Aftergood et al. [24], "cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction". Overall, cybersecurity typically concerns with the understanding of diverse cyber threats or attacks and corresponding defense strategies to prevent them, and eventually protect the systems, which is associated with confidentiality, integrity, and availability [50] [72] [103]. Based on these definitions, we can conclude that cybersecurity is all about the security of anything in the cyber realm, such as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 January 2021 doi:10.20944/preprints202101.0457.v1

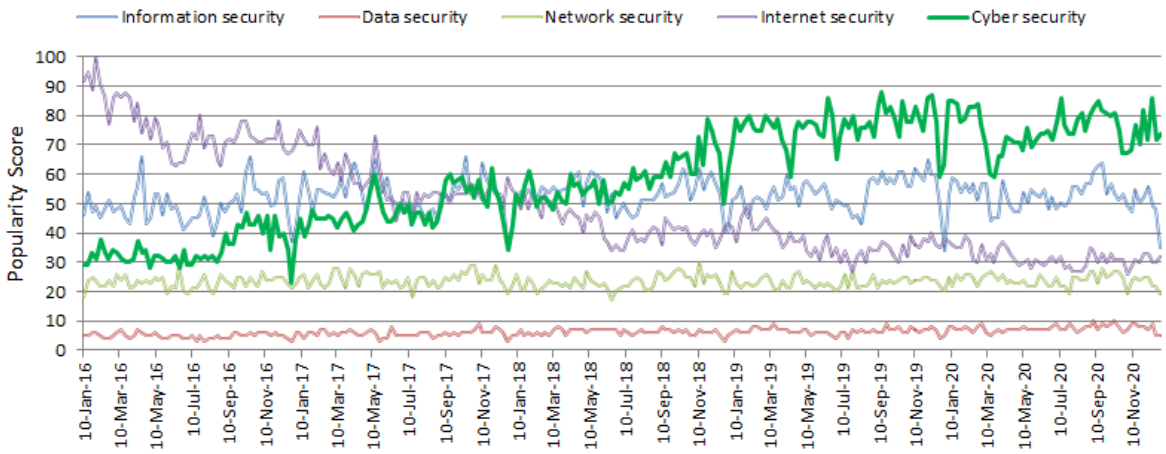

Fig. 1: The worldwide popularity score of cyber security comparing with relevant terms in a range of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score.

network security, information security, application security, operational security, Internet of Things (IoT) security, cloud security, infrastructure security, and relevant others. While traditional cybersecurity systems consist mainly of network protection systems and computer security systems [116], we aim to provide a wide range of cybersecurity view to the readers as it is one of the major concern in our digital life in various perspective, from commercial purpose to personalized mobile computing.

2.3 Security Incident and Attacks

A security incident is typically a malicious activity that threatens the security factors, i.e., confidentiality, integrity, and availability, defined earlier. Several types of cybersecurity incidents, i.e., cyber threats and attacks, may impact on an organization or an individual [154]. In general, a cyber-threat can be defined as a possible security violation that might exploit the vulnerability of a system or asset, while an attack is a deliberate unauthorized action on a system or asset. Cyber-attacks include threats like computer viruses, data breaches, denial of service (DoS) attacks, etc. In Table 1, we list the most common cyber-threats and attacks that are needed for consideration in today's cyber world.

2.4 Cybersecurity Defense Strategies

Cybersecurity defense strategies are typically for the protection of the computer systems and networks from the damage of the associated hardware, software, or data, as well as the disruption of the services they provide. More granularly, they are responsible for preventing data breaches or security incidents that can be defined as any kind of malicious or unauthorized activity to protect the systems [83]. In the following, we give an overview of traditional security mechanisms.

- *Access control* [123] is a security mechanism that typically regulates the access or use of the resources, e.g., computer networks, system files, or data, in a computing environment. For example, based on the responsibilities of individual users, an attribute or role-based access control scheme may be used to limit network access, reducing the risk to the company or entity.
- *Firewall* [178] is a security framework for the network that tracks and regulates incoming and outgoing network traffic. Firewalls are defined as a network-based or host-based system that is based on a set of security rules to allow or block the traffic. It is also capable of filtering traffic from unsecured or suspicious sources to avoid attacks, such as malicious traffic.
- *Anti-malware* [175] also known as antivirus software, is a computer program that is typically used to prevent, detect, and remove computer viruses, or malware. Modern antivirus software can protect users from various malware attacks such as ransomware, backdoors, trojan horses, worms, spyware, etc.
- *Sandbox* [69] is a security mechanism used for mitigating the system failures or software vulnerabilities from spreading through separating the running programs. It is often used to execute untrusted programs or code, possibly from unverified suppliers, users, websites, or untrusted third parties.
- *Security information and event management (SIEM)* [71] is a combination of security information man-

Table 1: The most common cyber-threats and attacks in cybersecurity

Key Terms	Description	References
Unauthorized access	an act of accessing information without authorization to the network, systems or data that results in a breach or violation of a security policy.	[154]
Malware	to cause extensive damage to data and systems or to obtain unauthorized access to a network, often referred to as malicious software or program.	[72]
Ransomware	a kind of malware attack that prevents users from accessing their device or personal files and needs a payment of ransom in order to regain access.	[108]
Backdoor	a type of malware attack that bypasses normal authentication or encryption to gain high-level user access to a computer device, network or software application.	[53] [166]
Malicious bot	a type of malware to steal information, or infect a host, often used by cyber criminals.	[72]
Typo-squatting Attacks	a form of cybersquatting, also known as URL hijacking or domain mimicry, fake URL, that tricks users into visiting a malicious website.	[42]
Denial of Service (DoS)	a type of cyber attack on a service that interferes with its normal functioning and prevents access to that by other users.	[72]
Distributed DoS (DDoS)	a large-scale DoS attack where the perpetrator uses multiple machines and networks.	[72]
Botnets	a collection of malware-infected internet-connected devices that allow hackers to carry out malicious activities such as leaks of credentials, unauthorized access, data theft and DDoS attacks.	[72]
Computer virus	a type of malicious software program loaded without the knowledge of the user onto a user's computer and performs malicious acts.	[72]
Social engineering	psychological manipulation of people that enable attackers to gain legitimate, authorized access to confidential information.	[72]
Phishing	a type of social engineering that involves fraudulent attempts to obtain sensitive information, such as details of banking and credit cards, login credentials, etc.	[72] [35]
Zero-day attack	is considered as the threat of an unknown security vulnerability.	[31] [44]
Cryptographic attack	to finding a weakness in a code, cipher, cryptographic protocol or key management scheme.	[111]
Insider threats	originates from within the organization by legitimate users, e.g., employees, to misuse access to networks and assets.	[168]
Supply chain attack	targets less secure supply network components to harm any industry, from the financial sector, oil or government sector.	[117] [56]
Man-in-the-middle (MiiM)	a type of cyberattack in which a malicious actor introduces himself into a two-party conversation to gain access to sensitive information.	[91]
Data Breaches	known as a data leakage, a theft of data by a malicious actor, e.g., unauthorized access of data by an individual, application, or service.	[148] [11]
Hacking	to compromise data and digital devices, such as computers, smartphones, tablets, and even entire networks.	[68] [35]
SQL injection attack	to execute malicious SQL statements for backend database manipulation to access information, typically used to attack data-driven applications.	[46]
Attacks on IoT Devices	to make it part of a DDoS attack and unauthorized access to data being collected by the device.	[141]
Malware on Mobile App	to get access of personal information, location data, financial accounts etc. by the malicious actor.	[160] [145]
Others	privilege escalation [55], password attack [77], advanced persistent threat [164], cryptojacking attack [150], web application attack [77], and so on.	

agement (SIM) and security event management (SEM) that provides real-time analysis of device and network hardware security alerts.
– *Cryptography* [23] is a popular method used for protecting data or information that uses the secret keys, e.g., secret-key, public key, and hash function, to encrypt and decrypt data for communication.

Although the traditional well-known security approaches have their own merits for different purposes, these might not be effective according to today's diverse needs in the cyber industry, because of lacking intelligence and dynamism [36] [112] [157] [159]. The intrusion detection system (IDS) becomes more popular that is typically defined as "a device or software application that monitors a computer network or systems for malicious activity or policy violations" [76]. IDS is typically capable to identify the diverse cyber threats and attacks, even the unknown zero-day attack, and able to respond in real-time based on the user's requirements. IDS gathers data from different sources in a computer

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 January 2021 doi:10.20944/preprints202101.0457.v1

network or device for this purpose and identifies security policy breaches that can be used to detect internal and external attacks [47] [124]. IDS can be several types based on environment type and detection approaches. For instance, based on the scope from single computers to large networks, the most common types of IDS are:

- *Host-based IDS (HIDS):* runs on a host, analyze traffic, and detect malicious or suspicious activity. Thus, it can provide real-time visibility into what's happening on the critical security systems, and which adds to the additional security [141].
- *Network-based IDS (NIDS):* On the other hand, NIDS analyzes and monitors network connections to detect malicious activity or policy violations on a network [141].

Similarly, IDS can be several types depending on the detection method, where the most well-known versions are the signature-based IDS and anomaly-based IDS [83].

- *Signature-based IDS (SIDS):* It looks for unique patterns, such as network traffic byte sequences, or recognized malicious sequences that the malware uses as signatures. It is also considered as misuse or knowledge-based detection that performs well for the known attacks [97]. It can, however, face the greatest challenge in detecting unknown or new attacks.
- *Anomaly-based IDS (AIDS):* On the other hand, due to the rapid growth of malware in recent days, AIDS is mainly used to detect unknown attacks. To detect anomalies like the unknown or zero-day attacks, machine learning techniques can also be used to build the protection model [30] [141].
- *Hybrid IDS:* The hybrid IDS is obtained by combining anomaly-based IDS with the misuse-based IDS discussed above and can be used to effectively detect the malicious activities in several cases [163] [174].
- *Stateful Protocol Analysis (SPA):* Besides, SPA is another type of method that identifies the deviations of protocol state. This approach is similar to the anomaly-based method, however it uses predetermined universal profiles of benign protocol activity [97].

Once the malicious activities have been detected, the intrusion prevention system (IPS) can be used to avoid and block them. This can be done in many ways, such as manual, sending notification, or automated operation [126]. Among these methods, an automated response system (ARS) may be more effective because it does not involve a human interface between the detection and response systems.

2.5 Cybersecurity Data and Systems

Research that relies on security information gathered from different sources is often problem-specific, which varies from application-to-application. A number of studies have been performed on cybersecurity systems and facilities that take into account different sources of security data. For instance, NSL-KDD [158] that contains security data related to various types of cyber-attacks such as denial of service (DoS), remote-to-local (R2L), user-to-remote (U2R), and probing attack. Another popular dataset UNSW-NB15 [115] that consists of different types of attacks. Similarly, several other datasets exist in the domain of cybersecurity, for instance, DARPA [174] [100], CAIDA [4] [3], ISOT'10 [16] [14], ISCX'12 [5] [149], CTU-13 [10], CIC-IDS [9], CIC-DDoS2019 [6], MAWI [73], ADFA IDS [173], CERT [99] [64], EnronSpam [13], SpamAssassin [19], LingSpam [17], DGA [1] [2] [12] [181], Malware Genome project [184], Virus Share [20], VirusTotal [21], Comodo [7], Contagio [8], DREBIN [92], Microsoft [18], Bot-IoT [89], etc. A summary of these cybersecurity datasets highlighting diverse attack-types and machine learning-based usage in different cyber applications are provided in our earlier paper Sarker et al. [141]. Several works focused on deep learning have recently been studied in the field. For example, methods of detection of network attacks based on deep learning techniques are studied in [171]. The researchers of [59] review deep learning for the detection of cyber security intrusion. In [32], the authors review deep learning-based intrusion detection systems. The authors of [43] conducted a study of cybersecurity deep learning methods. In [28], a survey of computer and deep learning techniques for internet of things (IoT) security is studied. We summarize several data-driven tasks and machine learning modeling used for various purposes in the cybersecurity domain in Table 2.

While different types of cybersecurity data and techniques mentioned above are used for various purposes in the field of cybersecurity and systems, there is an interest in security intelligence modeling in a broad sense, according to today's cyber industry needs. Therefore, in this paper, we intend to concentrate on a comprehensive view on "AI-driven cybersecurity" in terms of concepts and security modeling for intelligent cybersecurity services and management, where the most popular AI techniques such as machine and deep learning methods, the concept of natural language processing, knowledge representation and reasoning, as well as the concept of knowledge or rule-based expert systems modeling can be used. These AI methods based on security intelligence modeling can be used to solve

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 January 2021 doi:10.20944/preprints202101.0457.v1

Table 2: A summary of data-driven/machine learning tasks and approaches in the domain of cybersecurity.

Used Technique and Approaches	Purpose	References
Clustering	intrusion detection analysis	Chandrasekhar et al. [49], Sharifi et al. [147], Lin et al. [98]
Rule-based Approach	network intrusion detection systems	Tajbakhsh et al. [156], Mitchell et al. [110]
Support Vector Machines	attack classification intrusion detection and classification DDoS detection and analysis, anomaly detection systems	Kotpalliwar et al. [90], Pervez et al. [121] Yan et al. [176], Li et al. [96], Raman et al. [129], Kokila et al. [87], Xie et al. [172], Saxena et al. [144], Chandrasekhar et al. [49]
K-Nearest Neighbor	Network intrusion detection system reducing the false alarm rate intrusion detection system	Shapoorifard et al. [146], Vishwakarma et al. [165], Meng et al. [109], Dada et al. [52]
Naive Bayes	intrusion detection system	Koc et al. [86]
Decision Tree	malicious behavior analysis intrusion detection system anomaly detection system	Moon et al. [114], Ingre et al. [70], Malik et al. [106], Relan et al. [130], Rai et al. [127], Sarker et al. [134], Puthran et al. [122], Balogun et al. [41], Jo et al. [74]
Random Forests	network intrusion detection systems	Zhang et al. [183]
Adaptive boosting	network anomaly detection	Yuan et al. [179]
Neural Network and Deep Learning (RNN, LSTM, CNN)	anomaly intrusion detection attack classification Malware traffic classification	Jo et al. [74], Alrawashdeh et al. [34], Yin et al. [177], Kim et al. [85], Almiani et al. [33], Kolosnjaji et al. [88], Wang et al. [167]
Genetic Algorithm	preventing cyberterrorism and intrusion detection	Hansen et al. [67], Aslahi et al. [39], Azad et al. [40]
Hidden Markov Model	intrusion detection system	Ariu et al. [37], Aarnes et al. [38]
Reinforcement Learning	detecting malicious activities and intrusions	Alauthman et al. [29], Blanco et al. [45], Lopez et al. [102]

various cybersecurity issues and tasks, such as automatic identification of malicious activities, phishing detection, to detect malware, prediction of cyber-attacks, fraud detection, access control management, detection of anomalies or intrusions, etc. Thus, the concept of AI-based security intelligence modeling can enable the cybersecurity computing process to be more actionable and intelligent compared to conventional systems.

3 AI-based Security Intelligence Modeling

As discussed earlier, intelligent cybersecurity management is based on artificial intelligence, applies various AI methods that eventually seek for intelligent decision making in cyber applications or services. In our analysis, we have taken into account the most popular AI techniques that include Machine Learning (ML) and Deep Learning (DL) methods, the concept of Natural Language Processing (NLP), Knowledge Representation and Reasoning (KRR), as well as the concept of knowledge or rule-based Expert Systems (ES) modeling, according to today's need in the cyber industry. These AI methods based security intelligence modeling potentially can be used to make intelligent decisions in cybersecurity tasks, which are discussed briefly in the following.

3.1 Machine Learning based Modeling

Machine Learning (ML) including neural network-based deep learning is an important part of Artificial Intelligence (AI) that can be used to build effective security modeling utilizing the given historical cybersecurity data, summarized in Section 2. A security model for machine learning is typically a collection of target security-related data from different relevant sources, such as network behavior, database activity, application activity, or user activity, etc., and the algorithms chosen to operate on that data to deduce the performance [141]. In the following, we list several popular machine learning algorithms that can be used for different purposes ranging from exploiting malware to risky behavior identification that might lead to a phishing attack or malicious code within the area of cybersecurity.

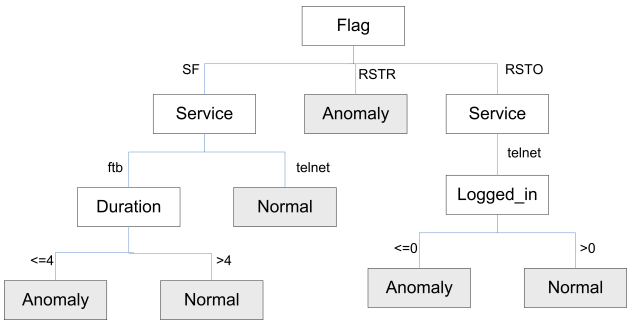

Fig. 2: An example of detecting cyber-anomalies based on a decision tree-based machine learning model.

– *Supervised learning:* Supervised learning is performed when specific target attack-anomaly classes are defined to reach from a certain set of inputs, i.e., task-driven approach [142]. For instance, to classify internal data, spam and malicious activities, supervised technique can be used. Navies Bayes [75], Various types of decision trees, such as C4.5 [125], IntrudTree [134], or even BehavDT [137] for behavioral pattern analysis, etc., can generate policy rules

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 January 2021 doi:10.20944/preprints202101.0457.v1

as well, K-nearest neighbors [27], Support vector machines [81], Adaptive boosting [62], Logistic regression [93], Stochastic Gradient Descent [65], or ensemble methods such as XGBoost [118], Random Forest learning [48], etc. are the well-known classification techniques in the area. These techniques can be used for data-driven security modeling according to their learning capabilities from the security data, e.g., classifying and predicting malware attacks or cyber-anomalies. For instance, a decision tree-based machine learning model, e.g., IntruDTree model [134], to detecting cyber-anomalies, is shown in Figure 2, which provides a significant accuracy 98% for unseen test cases.

– *Unsupervised learning:* Security data is not labeled or categorized always in the real world scenario. Thus unsupervised learning, i.e., data-driven approach, can be used to find patterns, structures, or knowledge from unlabeled data [142]. The hidden patterns and structures of the datasets can be uncovered by clustering, a common form of unsupervised learning. Clustering techniques can group the security data by taking into account certain measures of similarity in the data. Several clustering algorithms, for example, partitioning methods such as K-means [104], K-medoids [131], CLARA [80], etc., density-based methods such as DBSCAN [58], distribution-based clustering such as Gaussian mixture models (GMMs) [118], hierarchical-based methods, agglomerative or divisive such as Single linkage [151], Complete linkage [152], BOTS [138], etc. can be used in such purposes. Moreover, incident response and risk management from recommendation methods is another area that typically comes from association learning techniques. Several methods such as AIS [25], Apriori [26], FP-Tree [66], RARM [54], Eclat [182], ABC-RuleMiner [140] can be used for building rule-based machine learning model, e.g., policy-rule generation.
– *Security feature optimization :* Today's cybersecurity datasets may contain security features with high dimensions [134]. Thus, to minimize the complexity of a security model, feature optimization is important. Therefore the task of feature selection or feature engineering such as considering a subset of security features according to their importance or significance in modeling, the extraction of features considering the key components, or generating new features could help simplify as well as optimize the resultant security model. Several methods such as variance threshold [118], Pearson's correlation coefficient defined for two variables (X and Y) in Equ. 1 [65], analysis of variance (ANOVA) [118], chi-squared test considering O_i as observed value and E_i as expected value in Equ. 2 [118], recursive feature elimination (RFE) [118], principal component analysis (PCA) [135], or model-based selection [118] [134] etc. can be used to perform the tasks according to the characteristics or nature of the security data. For example, the authors take into account the ranking of security features in [134], according to their significance, in order to create an efficient tree-based security model that achieves 98% with the simplified model for unseen test cases.

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (1)$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

– *Deep learning and others :* Deep learning is typically considered as part of a broader family of machine learning approaches, originating from an Artificial Neural Network (ANN). In Fig. 3, we show a structure of artificial neural network modeling considering input, hidden, and output layer, for detecting cyber anomalies or attacks. In the domain of cybersecurity, the deep learning methods can be used for various purposes such as detecting network intrusions, detecting and classifying malware traffic, backdoor attacks, etc. [174] [166] [59]. Multi-layer perceptron (MLP) [161], convolutional neural network (CNN) [101], recurrent neural network (RNN) and long-short term memory (LSTM) are the popular approaches used in deep learning modeling [53] [101] [177]. In these deep learning models, many hidden layers can be used to complete the overall computing process. The strongest aspect of deep learning techniques is effectively learning feature hierarchies based on the patterns in the data [32]. Several unsupervised techniques such as autoencoder (AE), deep belief network (DBN), restricted Boltzmann machines (RBMs), generative adversarial network (GAN) etc., can also be used in the domain of cybersecurity [171] [32]. Hybrid techniques can also be used for significant outcomes in several cases [32]. For instance, an intrusion detection model based on the LSTM architecture with RNN achieved an attack detection percentage of 98.8% [85]. A deep learning model based on a stacked auto-encoder with a soft-max classifier for efficient network intrusion detection is proposed in [82], which achieves up to 99.99% accuracy for the KDD99 dataset, and 89.13% for the UNSW-NB15 dataset. Besides the semi-supervised

learning combining the supervised and unsupervised techniques discussed above, and reinforcement learning techniques such as Monte Carlo learning, Q-learning, Deep Q Networks [78] [141] can be used in the area.

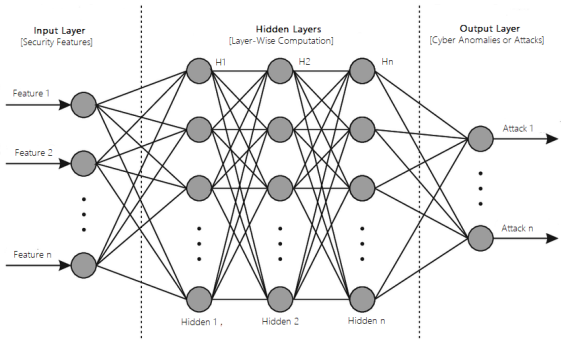

Fig. 3: A structure of artificial neural network modeling for detecting cyber anomalies or attacks with multiple processing layers.

Thus the machine and deep learning methods discussed above can play a vital role to understand and analyze the actual phenomena with cybersecurity data, depending on the nature or characteristics of the security features and the sufficient amount of data needed for learning. These techniques can extract insights or useful knowledge from the given security data and eventually build a data-driven security model. Such models can learn from the training data and behave accordingly for the unseen test cases. Overall, the resultant machine learning-based security models can make intelligent cybersecurity decisions through analyzing data from the huge amount of cyber events. Therefore, we can conclude that machine learning security models would be able to alter the future of cybersecurity applications and industry, because of their data learning capabilities, and could be a major part in the domain of AI-driven cybersecurity.

3.2 NLP-based Modeling

Natural Language Processing (NLP) is considered as an important branch of AI that can make it possible for computers to understand human language, interpret it, and eventually determine which parts are important in an intelligent system [139]. NLP is increasingly used nowadays by cybercriminals and security defense tools in the understanding and processing of unstructured data generated. NLP's ultimate aim is to extract knowledge from unstructured data or information, i.e., to interpret, decipher, comprehend, and make sense of human languages in a valuable way. In the following, we discuss several parts of NLP that can be used for intelligent cybersecurity modeling when unstructured security content is available.

- *Lexical analysis:* It usually includes the arrangement of terms being described and analyzed. Lexical analysis separates the entire chunk of text according to the criteria into paragraphs, sentences, phrases, or tokens such as identifier, keyword, literal, etc. For example, the lexical analysis of domain names [84] will lead to the development of the NLP-based model to classify the malicious domains that may encompass the "malicious nature" of the domains used by cybercriminals.
- *Syntactic analysis:* This is seen as one of the key tools used to complete the tasks of the NLP, which is used to determine how the natural language aligns with the grammatical rules. The most widely used techniques in NLP are: lemmatization, morphological segmentation, word segmentation, part-of-speech marking, parsing, sentence breaking, stemming, etc. A syntactic analysis, e.g., parsing [120], may contribute to developing an NLP-based model for cyberattack prediction, for example, to quickly extract useful data from large quantities of public text.
- *Semantic analysis:* Another of the key methods used to complete NLP assignments is semantic analysis, which includes understanding the context and perception of words and how sentences are structured. For example, for phishing classification, latent semantic analysis can be used with keyword extraction [94]. The most widely used techniques in NLP are entity recognition (NER), word sense disambiguation, natural language generation, etc. For example, a NER-based automated system [63], can be used to diagnose cybersecurity situations in IoT networks.

Several most frequently used algorithms such as Bag-of-Words (BoW), TF-IDF (term frequency-inverse document frequency), Tokenization and Stop Words Removal, Stemming, Lemmatization, Topic Modeling, etc. are used in the area of NLP [155]. Most of the NLP-based modeling relies on machine and deep learning techniques discussed above for building the resultant data-driven model that can be used for various purposes in the domain of cybersecurity. In the following, we give examples of NLP-based security modeling.

- *Detecting malicious domain names* – to identify malicious domain names (e.g., clbwpvdyztoepfua.lu) from benign domains (e.g., cnn.com), the NLP methods can be used. It helps to build a technique for detecting such malicious domains in DNS traffic based on

the patterns that are inherent in domain names using a domain dataset collected via a domain crawl.
- *Vulnerability analysis* – to detect the weaknesses and vulnerabilities in the code, the NLP techniques can be used. For instance, n-grams and various smoothing algorithms [113] combined with machine learning can be used to build such a model based on the associated patterns for detecting vulnerabilities. One example could be the detection of zero-day vulnerabilities in the banking sector. The analysts usually study conversations on various platforms on the web and looking for the relevant information that is useful for the purposes.
- *Phishing identification* – detection of a phishing attack is a challenging problem, because of considering this as semantics-based attacks. Phishing can be several categories, such as web page based, email content-based, URL based, etc. A machine learning model with a set of features can be used to detect such phishing [57]. NLP techniques can be used to effectively extract the features from such content as well as to build the model.
- *Malware family analysis* – to modeling behavioral reports into a series of words is necessary to effectively detect malware. For the formulation of behavioral reports [79], a bag-of-words (BoW) NLP model might be helpful. For the automated engineering of related security features and to construct the model, NLP with machine learning techniques can be used.

Overall, to enhance the cybersecurity operations by automating threat intelligence extracted from the unstructured sources, an NLP-based methodology can be used. Thus, NLP with the machine learning techniques is considered as the driver for the automation of security activities according to its capabilities in security modeling depending on the target security application. Therefore, we can conclude that NLP-based security modeling could be another major part of the domain of AI-driven cybersecurity.

3.3 Knowledge Representation and Conceptual Modeling

Knowledge representation and reasoning is another field of AI that typically represents the real-world information so that an intelligent cybersecurity system can utilize that information to solve complex security problems like a human. In the real world, knowledge of cybersecurity is usually regarded as information about a specific security domain. It is the analysis of how an intelligent cybersecurity agent's views, intentions, and decisions can be adequately articulated for automated reasoning, e.g., inference engines, classifiers, etc., to solve complex security problems. In this section, we first discuss and summarize the approaches of knowledge representation, and then we discuss a conceptual security model based on knowledge.

3.3.1 Knowledge Representation

Modeling the intelligent actions of a security agent is the key purpose of knowledge representation. In the field of cybersecurity, it enables a computer to benefit from that knowledge of security and function like a human being accordingly. Instead of considering the bottom-up learning, it takes into account a top-down approach to build the model to behave intelligently. As discussed in [139], descriptive knowledge, structural knowledge, procedural knowledge, meta knowledge, heuristic knowledge, etc. are the several types of knowledge that can be used in various application areas. In the following, we summarize several knowledge representation methods such as logical, semantic network, frame, and production rules [153], that can be used to build a knowledge-based conceptual model.

- *Logical representation:* It represents with concrete rules without any ambiguity that typically deals with propositions. Thus, logic can be used to represent simple facts that are the general statements that may be either 'True' or 'False'. Overall, logical representation means drawing a conclusion based on various conditions. Although logical representation enables us to do logical reasoning, the inference may not be so efficient due to the restrictions and challenges to work with.
- *Semantic network representation:* We may represent our information in the form of graphical networks within semantic networks. This network is made up of objects and arcs representing nodes that define the relationship between those objects. Overall, they provide a structural representation of statements about a domain of interest. Although semantic networks are a natural representation of information, their intelligence in action depends on the system's creator.
- *Frame representation:* A frame, derived from semantic networks, is a structure-like record that consists of a set of attributes to represent an object in the world and its values. In the frame, knowledge about an object or event can be stored together in the knowledge base. Although frame representation is easy to understand and visualize, it cannot proceed with the inference mechanism smoothly.
- *Production rules:* It typically consists of pairs of the condition, and corresponding action, which means,

"If condition then action". Thus an agent first checks the condition and then the corresponding rule fires if the condition satisfies. The main advantage of such a rule-based system in cybersecurity is that the "condition" part can determine which rule is suitable to apply for a specific security problem. And the "action" part carries out the solutions associated with that problem. Thus in a rule-based cybersecurity system, it allows us to remove, add or modify the rules according to the needs.

Overall, we can say that the knowledge for building a knowledge-based conceptual model or system can be represented in multiple ways. However, the effectiveness of these methods in a security system may vary depending on the nature of the data and target application. In the following, we discuss how security ontologies, a formal way to define the semantics of knowledge and data, can be used to build a conceptual security model.

3.3.2 Security Ontologies and Conceptual Modeling

Ontologies, through information representation techniques, are conceptual models of what exists in some domain, brought into machine-interpretable form. Top-level ontologies or upper ontologies, domain ontologies, and application ontologies are several types of ontologies used in the area [153]. In general, ontology is "an explicit specification of conceptualization and a formal way to define the semantics of knowledge and data" [105]. According to [105], formally, an ontology is represented as "$\{O = C, R, I, H, A\}$, where $\{C = C_1, C_2, ..., C_n\}$ represents a set of concepts, and $\{R = R_1, R_2, ..., R_m\}$ represents a set of relations defined over the concepts. I represents a set of instances of concepts, and H represents a Directed Acyclic Graph (DAG) defined by the subsumption relation between concepts, and A represents a set of axioms bringing additional constraints on the ontology". In an ontology-based information security, five concepts such as threat, vulnerability, attack, impact, and control, might be involved [119].

- *Concept:Threat* represents various types of difficulties or dangers against a given set of security properties.
- *Concept: Vulnerability* mainly represents the weaknesses of a cybersecurity system.
- *Concept:Attack* represents various types of security incidents caused by cyber criminals.
- *Concept:Impact* represents the effects that a security incident can imply.
- *Concept: Controls* represents the relevant mechanisms that can be used to reduce or avoid the effects of a security incident or to protect a vulnerability.

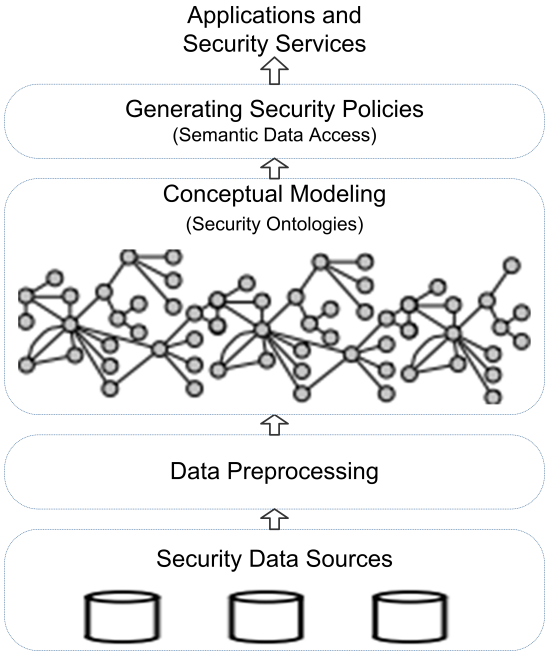

Fig. 4: A structure of conceptual modeling based on security ontologies in a cybersecurity system and the corresponding information flow from data source to application.

Based on these concepts and their relationships, a conceptual security model can be built to solve complex security problems. The rationale behind the conceptual security model can be structured as: a cyber-threat may produce an attack or security incident that exploits the vulnerabilities of the system, which may have an impact on that system. A control mechanism that can detect, prevent, or block the attack, is thus needed to protect the system and make it secured. In Fig. 4, we show a structure of conceptual modeling based on security ontologies in a cybersecurity system and the corresponding information flow from data source to application. According to Fig. 4, the automated security policies can also be generated from the relevant security ontologies that are used in the eventual security services or applications. Thus it is capable of making intelligent decisions according to the concepts and their semantic relationships that exist in the ontologies. Based on different knowledge representation formalisms, various ontology languages can be used. In the area of semantic web, Web Ontology Language (OWL) [107] is mostly used to formalize and represent these concepts and their semantic relationships in a graphical representation to build an ontology-based security model. Overall, we can conclude that knowledge representation based conceptual security modeling could be another part in the domain of AI-driven cybersecurity according to its computing capabilities while making intelligent decisions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 January 2021 doi:10.20944/preprints202101.0457.v1

3.4 Cybersecurity Expert System Modeling

In artificial intelligence, an expert system is generally a computer system that emulates the decision-making capacity of a human expert. A cybersecurity expert system is an instance of a knowledge-based or rule-based system in which decisions can be made based on security guidelines. The system is typically split into two subsystems, such as the inference engine and the knowledge base represented as security rules, as shown in fig. 5.

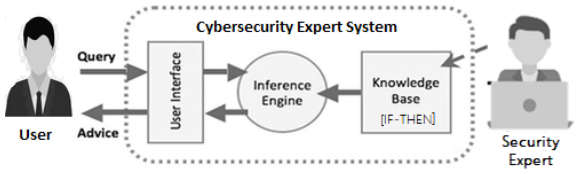

Fig. 5: A structure of a cybersecurity expert system modeling.

The foundation of this cybersecurity expert framework is the knowledge base shown in Fig. 5, as it consists of knowledge of the domain of the target cybersecurity application as well as operational knowledge of the rules of security decisions. The inference engine shown in Fig. 5, on the other hand, applies the rules to known facts from a security perspective to deduce new facts. The user interface shown in Fig. 5 recognizes the original security facts and invokes the inference engine to trigger the knowledge base decision rules.

Usually, a rule consists of two parts: the antecedent (IF part), called the state or premise, and the inference or action called the consequent (THEN part). Thus, a rule's basic syntax can be expressed as:
IF $<$ *antecedent* $>$ THEN $<$ *consequent* $>$

For instance, "if the flag value is $RSTR$, then the outcome is *anomaly*" can be an example of the IF-THEN rule for detecting anomalies. Similarly, another rule with multiple security features could be "if flag value is SF, service is ftb, and duration $<= 4$, then the outcome is *anomaly*", generated from the tree shown in Fig. 2. In addition to human experts, several techniques can be used to generate rules that can be used to build the rule-based cybersecurity expert system.

- *Classification learning rules:* In machine learning, the classification is one of the popular techniques that can be used in various application areas. Several popular classification techniques such as decision trees [125], IntrudTree [134], BehavDT [137], Ripple Down Rule learner (RIDOR) [169], Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [170], etc. exist with the ability of rule generation.
- *Association learning rules:* In general association rules are created by searching for frequent IF-THEN pattern data on the basis of [140] support and confidence value. For generating rules using a given data set, common association rule learning techniques such as AIS [25], Apriori [26], FP-Tree [66], RARM [54], Eclat [182], ABC-RuleMiner [140], etc. can be used.
- *Fuzzy logic-based rules:* Usually, fuzzy logic is an approach to computing focused on "degrees of truth" rather than the usual "true or false" (1 or 0) [180]. Thus, instead of Boolean logic, a fuzzy rule-based expert system uses fuzzy logic. In other words, by using these rules, a fuzzy expert system is a set of membership functions and rules that can provide outputs.
- *Conceptual semantic rule:* As discussed earlier, an ontology is "an explicit specification of conceptualization and a formal way to define the semantics of knowledge and data" [105]. For instance, security ontologies include the relationships between each entry within an ontology that can be used to generate such conceptual rules. As each security decision must consider the concrete company environment, particular domain ontology can help for building an effective semantic cybersecurity application.

Thus, a rule-based cybersecurity expert system model may have the decision-making capacity of a security expert in an intelligent cybersecurity framework that is built to solve complex cybersecurity issues, as well as by information reasoning. A rule generation method discussed above can play a major role in generating the IF-THEN rules while developing the knowledge base module. The rules can then be modified and handled according to the requirements by domain experts with knowledge of business rules. Overall, we can conclude that cybersecurity expert systems modeling could be another important part in the domain of AI-driven cybersecurity according to its computing capabilities while making intelligent decisions.

4 Research Issues and Future Directions

As we have discussed the role of Artificial Intelligence (AI) throughout the paper, which is known as the key technologies of the Fourth Industrial Revolution (Industry 4.0), can play a significant role for intelligent cybersecurity services and management. To intelligently solve today's various cybersecurity issues, i.e., protecting of Internet-connected systems from cyber-threats,

attacks, damage, or unauthorized access, popular AI methods such as machine and deep learning, natural language processing, knowledge representation and reasoning, as well as the concept of knowledge or rule-based expert systems modeling can be used, discussed briefly in Section 3. However, several research issues that are identified within the area of AI-driven cybersecurity, discussed briefly in the following.

According to our study in this paper, cybersecurity source datasets are the primary component, especially to extract security insight or useful knowledge from security data using machine and deep learning technique, discussed briefly in Section 3. Thus, the primary and most fundamental challenge is to understand the real-world security issues and to explore the relevant cybersecurity data to extract insights or useful knowledge for future actions. For instance, public text data such as cyber-related webpage text is used to detect and track the potential cyberattacks [120]. However, collecting the security data is not straight forward as the data sources could be multiple and dynamic. Thus, collecting various types of real-world data such as structured, semi-structured, unstructured, etc. relevant to a particular problem domain with legal access, which may vary from application to application, is challenging. Therefore, to understand the security problem, and to integrate and manage the collected data for effective data analysis could be one of the major challenges to work in the area of AI-driven cybersecurity.

The next challenge could be an effective and intelligent solution to tackle the target security problems. Although several machine and deep learning techniques, such as clustering, rule-based approach, classification, neural network, etc. [141] are employed to solve several security problems, summarized in Table 2, these models can be improved with advanced analytics. For instance, observing attack patterns in time-series, behavioral analysis, data sparseness in security analysis, the impact of security features in modeling, simplifying and optimizing the security model, taking into account advanced feature engineering tasks, synchronizing temporal patterns in modeling while considering multiple data sources, etc. can be considered. Moreover, several important issues such as data aggregation, redundancy in rule generation, effectiveness of prediction algorithms, data inconsistency, recent pattern analysis for prediction [132] [133] [136], etc. might be an important issue for effective data-driven modeling. Thus, advanced analytics techniques, improved machine or deep learning techniques, new data-driven algorithms, or hybrid methods could give better results for modeling security intelligence, depending on the nature of the security problems, which could be a potential research direction in the area.

Besides, to effectively extract the useful insights from the unstructured security data and to effectively build an intelligent security model could be another issue. For instance, a large amount of textual content is needed to analyze identifying malicious domains, security incident and event management, malware family analysis, domain classification, phishing, source code vulnerability analysis, spam emails, etc., that are discussed briefly in Section 3. Therefore effectively mining the relevant contents using natural language processing (NLP) techniques, or designing a new NLP based model, could be another research direction in the area of AI-driven cybersecurity. An effective cybersecurity expert system modeling considering IF-THEN policy rules could be another potential research direction in the area. However, the development of large-scale rule-based systems in the area of cybersecurity may face numerous challenges. For instance, the reasoning process in the expert system can be very complex, difficult to manage [139]. Thus, a lightweight rule-based inference engine that allows to reason for intelligent cybersecurity services is important. Although several rule mining techniques are popular in the area, mentioned in Section 3, a concise set of security policy rules considering generalization, reliability, non-redundancy, exceptional discovery, etc., could make the expert security system more effective. Therefore, a deeper understanding and designing an effective rule-based system by taking into these properties could be another research issue in the area of AI-driven cybersecurity. Moreover, designing security ontologies according to today's need, or knowledge representation model, and eventually to build an effective conceptual security modeling, could be another potential research scope in the area.

Overall, the most important task for an intelligent cybersecurity system is to design and build an effective cybersecurity framework that supports the artificial intelligence techniques, discussed in Section 3. In such a framework, we need to take into account AI-based advanced analytics, so that the security framework is capable to resolve the associated issues intelligently. Therefore, to assess the feasibility and effectiveness of the related AI-based approaches, a well-designed cybersecurity framework and experimental evaluation are required, which is a very important direction and a major challenge as well. Overall, we can conclude that this paper has uncovered lots of research issues and potential future directions to resolve, discussed above, in the area of AI-driven cybersecurity.

5 Conclusion

Motivated by the growing significance of cybersecurity and artificial intelligence, in this paper, we have studied AI-driven cybersecurity. Our goal was to provide a comprehensive overview of how artificial intelligence can play a significant role in intelligent decision making and to build smart and automated cybersecurity systems. For this, we have presented security intelligence modeling where various AI-based methods such as machine and deep learning, the concept of natural language processing, knowledge representation and reasoning, as well as the concept of knowledge or rule-based expert systems modeling are used to intelligently tackle the cybersecurity issues. Such AI-based modeling can be used in various problem domains ranging from malware analysis to risky behavior identification that might lead to a phishing attack or malicious code, which are discussed briefly throughout this paper.

In the field of AI-driven cybersecurity, the concept of AI-based security intelligence modeling discussed in this paper can help the cybersecurity computing process to be more actionable and intelligent. Based on our study, we have also highlighted several research issues and potential directions that can help researchers do future research in the area. Overall, we believe this paper can be served as a reference point and guidelines for cybersecurity researchers as well as industry professionals in the area, especially from an AI-based technical point of view.

Compliance with ethical standards

Conflict of interest The author declares no conflict of interests.

References

1. Alexa top sites. available online: https://aws.amazon.com/alexa-top-sites/ (accessed on 20 october 2019).
2. Bambenek consulting—master feeds. available online: http://osint.bambenekconsulting.com/feeds/ (accessed on 20 october 2019).
3. Caida anonymized internet traces 2008 dataset. http://www. caida.org/data/passive/passive-2008-dataset.xml/ (accessed on 20 october 2019).
4. Caida ddos attack 2007 dataset. http://www.caida.org/data/ passive/ddos-20070804-dataset.xml/ (accessed on 20 october 2019).
5. Canadian institute of cybersecurity, university of new brunswick, iscx dataset, url http://www.unb.ca/cic/datasets/index.html/ (accessed on 20 october 2019).
6. Cic-ddos2019 [online]. available: https://www.unb.ca/cic/datasets/ddos-2019.html/ (accessed on 28 march 2020).
7. Comodo. available online: https://www.comodo.com/home/internet-security/updates/vdp/database. php (accessed on 20 october 2019).
8. Contagio. available online: http://contagiodump.blogspot.com/ (accessed on 20 october 2019).
9. Cse-cic-ids2018 [online]. available: https://www.unb.ca/cic/ datasets/ids-2018.html/ (accessed on 20 october 2019).
10. The ctu-13 dataset. available online: https://stratosphereips.org/category/datasets-ctu13 (accessed on 20 october 2019).
11. Data breach investigations report 2019, https://enterprise.verizon.com/resources/reports/dbir/ (accessed on 20 october 2019).
12. Dgarchive. available online: https://dgarchive.caad.fkie.fraunhofer.de/site/ (accessed on 20 october 2019).
13. Enronspam. available online: https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/enron-spam/ (accessed on 20 october 2019).
14. The honeynet project. http://www.honeynet.org/chapters/france/ (accessed on 20 october 2019).
15. Ibm security report, https://www.ibm.com/security/data-breach (accessed on 20 october 2019).
16. Isot botnet dataset. https://www.uvic.ca/engineering/ece/isot/datasets/index.php/ (accessed on 20 october 2019).
17. Lingspam. available online: https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/lingspampublic.tar.gz/ (accessed on 20 october 2019).
18. Microsoft malware classification (big 2015). available online: http://arxiv.org/abs/1802.10135/ (accessed on 20 october 2019).
19. Spamassassin. available online: http://www.spamassassin.org/publiccorpus/ (accessed on 20 october 2019).
20. Virusshare. available online: http://virusshare.com/ (accessed on 20 october 2019).
21. Virustotal. available online: https://virustotal.com/ (accessed on 20 october 2019).
22. Google trends. In *https://trends.google.com/trends/*, 2019.
23. Omar G Abood and Shawkat K Guirguis. A survey on cryptography algorithms. *International Journal of Scientific and Research Publications*, 8(7):410–415, 2018.
24. Steven Aftergood. Cybersecurity: The cold war online. *Nature*, 547(7661):30, 2017.
25. Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM, 1993.
26. Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
27. David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.
28. Mohammed Ali Al-Garadi, Amr Mohamed, Abdulla Al-Ali, Xiaojiang Du, Ihsan Ali, and Mohsen Guizani. A

survey of machine and deep learning methods for internet of things (iot) security. *IEEE Communications Surveys & Tutorials*, 2020.
29. Mohammad Alauthman, Nauman Aslam, Mouhammd Al-kasassbeh, Suleman Khan, Ahmad Al-Qerem, and Kim-Kwang Raymond Choo. An efficient reinforcement learning-based botnet detection approach. *Journal of Network and Computer Applications*, 150:102479, 2020.
30. Ammar Alazab, Michael Hobbs, Jemal Abawajy, and Moutaz Alazab. Using feature selection for intrusion detection system. In *2012 International Symposium on Communications and Information Technologies (ISCIT)*, pages 296–301. IEEE, 2012.
31. Mamoun Alazab, Sitalakshmi Venkatraman, Paul Watters, Moutaz Alazab, et al. Zero-day malware detection based on supervised learning algorithms of api call signatures. 2010.
32. AM Aleesa, BB Zaidan, AA Zaidan, and Nan M Sahar. Review of intrusion detection systems based on deep learning techniques: coherent taxonomy, challenges, motivations, recommendations, substantial analysis and future directions. *Neural Computing and Applications*, 32(14):9827–9858, 2020.
33. Muder Almiani, Alia AbuGhazleh, Amer Al-Rahayfeh, Saleh Atiewi, and Abdul Razaque. Deep recurrent neural network for iot intrusion detection system. *Simulation Modelling Practice and Theory*, page 102031, 2019.
34. Khaled Alrawashdeh and Carla Purdy. Toward an online anomaly intrusion detection system based on deep learning. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 195–200. IEEE, 2016.
35. A Alsayed and A Bilgrami. E-banking security: Internet hacking, phishing attacks, analysis and prevention of fraudulent activities. *Int. J. Of Emerg. Techn. and Adv. Activ*, 7(1):109–115, 2017.
36. Shahid Anwar, Jasni Mohamad Zain, Mohamad Fadli Zolkipli, Zakira Inayat, Suleman Khan, Bokolo Anthony, and Victor Chang. From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions. *Algorithms*, 10(2):39, 2017.
37. Davide Ariu, Roberto Tronci, and Giorgio Giacinto. Hmmpayl: An intrusion detection system based on hidden markov models. *computers & security*, 30(4):221–241, 2011.
38. André Årnes, Fredrik Valeur, Giovanni Vigna, and Richard A Kemmerer. Using hidden markov models to evaluate the risks of intrusions. In *International Workshop on Recent Advances in Intrusion Detection*, pages 145–164. Springer, 2006.
39. BM Aslahi-Shahri, Rasoul Rahmani, M Chizari, A Maralani, M Eslami, Mohammad Javad Golkar, and A Ebrahimi. A hybrid method consisting of ga and svm for intrusion detection system. *Neural computing and applications*, 27(6):1669–1676, 2016.
40. Chandrashekhar Azad and Vijay Kumar Jha. Genetic algorithm to solve the problem of small disjunct in the decision tree based intrusion detection system. *International Journal of Computer Network and Information Security (IJCNIS)*, 7(8):56, 2015.
41. Abdullateef Oluwagbemiga Balogun and Rasheed Gbenga Jimoh. Anomaly intrusion detection using an hybrid of decision tree and k-nearest neighbor. 2015.
42. Anirban Banerjee, Md Sazzadur Rahman, and Michalis Faloutsos. Sut: Quantifying and mitigating url typosquatting. *Computer Networks*, 55(13):3001–3014, 2011.
43. Daniel S Berman, Anna L Buczak, Jeffrey S Chavis, and Cherita L Corbett. A survey of deep learning methods for cyber security. *Information*, 10(4):122, 2019.
44. Leyla Bilge and Tudor Dumitraş. Before we knew it: an empirical study of zero-day attacks in the real world. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 833–844. ACM, 2012.
45. Roberto Blanco, Juan J Cilla, Samira Briongos, Pedro Malagón, and José M Moya. Applying cost-sensitive classifiers with reinforcement learning to ids. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 531–538. Springer, 2018.
46. Stephen W Boyd and Angelos D Keromytis. Sqlrand: Preventing sql injection attacks. In *International Conference on Applied Cryptography and Network Security*, pages 292–302. Springer, 2004.
47. Imen Brahmi, Hanen Brahmi, and Sadok Ben Yahia. A multi-agents intrusion detection system using ontology and clustering techniques. In *IFIP International Conference on Computer Science and its Applications*, pages 381–393. Springer, 2015.
48. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
49. AM Chandrasekhar and K Raghuveer. Confederation of fcm clustering, ann and svm techniques to implement hybrid nids using corrected kdd cup 99 dataset. In *2014 International Conference on Communication and Signal Processing*, pages 672–676. IEEE, 2014.
50. National Research Council et al. Toward a safer and more secure cyberspace. 2007.
51. D Craigen, N Diakun-Thibault, and R Purse. Defining cybersecurity. technology innovation management review, 4 (10), 13-21, 2014.
52. EG Dada. A hybridized svm-knn-pdapso approach to intrusion detection system. In *Proc. Fac. Seminar Ser.*, pages 14–21, 2017.
53. Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.
54. Amitabha Das, Wee-Keong Ng, and Yew-Kwong Woon. Rapid association rule mining. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 474–481. ACM, 2001.
55. Lucas Davi, Alexandra Dmitrienko, Ahmad-Reza Sadeghi, and Marcel Winandy. Privilege escalation attacks on android. In *international conference on Information security*, pages 346–360. Springer, 2010.
56. Shannon Eggers. A novel approach for analyzing the nuclear supply chain cyber-attack surface. *Nuclear Engineering and Technology*, 2020.
57. Gal Egozi and Rakesh Verma. Phishing email detection using robust nlp techniques. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 7–12. IEEE, 2018.
58. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
59. Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, and Helge Janicke. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50:102419, 2020.

60. Eric A Fischer. Cybersecurity issues and challenges: In brief, 2014.
61. Farhad Foroughi and Peter Luksch. Data science methodology for cybersecurity projects. *arXiv preprint arXiv:1803.04219*, 2018.
62. Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer, 1996.
63. Tiberiu-Marian Georgescu, Bogdan Iancu, and Madalina Zurini. Named-entity-recognition-based automated system for diagnosing cybersecurity situations in iot networks. *Sensors*, 19(15):3380, 2019.
64. Joshua Glasser and Brian Lindauer. Bridging the gap: A pragmatic approach to generating insider threat data. In *2013 IEEE Security and Privacy Workshops*, pages 98–104. IEEE, 2013.
65. Jiawei Han, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. 2011.
66. Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM Sigmod Record*, volume 29, pages 1–12. ACM, 2000.
67. James V Hansen, Paul Benjamin Lowry, Rayman D Meservy, and Daniel M McDonald. Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection. *Decision Support Systems*, 43(4):1362–1374, 2007.
68. Sunghyuck Hong. Survey on analysis and countermeasure for hacking attacks to cryptocurrency exchange. *Journal of the Korea Convergence Society*, 10(10):1–6, 2019.
69. Tyler Hunt, Zhiting Zhu, Yuanzhong Xu, Simon Peter, and Emmett Witchel. Ryoan: A distributed sandbox for untrusted computation on secret data. *ACM Transactions on Computer Systems (TOCS)*, 35(4):1–32, 2018.
70. Bhupendra Ingre, Anamika Yadav, and Atul Kumar Soni. Decision tree based intrusion detection system for nsl-kdd dataset. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 207–218. Springer, 2017.
71. Muhammad Irfan, Haider Abbas, Yunchuan Sun, Anam Sajid, and Maruf Pasha. A framework for cloud forensics evidence collection and analysis using security information and event management. *Security and Communication Networks*, 9(16):3790–3807, 2016.
72. Julian Jang-Jaccard and Surya Nepal. A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences*, 80(5):973–993, 2014.
73. Xuyang Jing, Zheng Yan, Xueqin Jiang, and Witold Pedrycz. Network traffic fusion and analysis against ddos flooding attacks with a novel reversible sketch. *Information Fusion*, 51:100–113, 2019.
74. Seongrae Jo, Haengnam Sung, and Byunghyuk Ahn. A comparative study on the performance of intrusion detection using decision tree and artificial neural network models. *Journal of the Korea Society of Digital Industry and Information Management*, 11(4):33–45, 2015.
75. George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
76. Leighton Johnson. Computer incident response and forensics team management: Conducting a successful incident response. 2013.
77. Bojan Jovičić and Dejan Simić. Common web application attack types and security using asp .net. *ComSIS. December*, 2006.
78. Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
79. ElMouatez Billah Karbab and Mourad Debbabi. Maldy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports. *Digital Investigation*, 28:S77–S87, 2019.
80. Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.
81. S. Sathiya Keerthi, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural computation*, 13(3):637–649, 2001.
82. Farrukh Aslam Khan, Abdu Gumaei, Abdelouahid Derhab, and Amir Hussain. A novel two-stage deep learning model for efficient network intrusion detection. *IEEE Access*, 7:30373–30385, 2019.
83. Ansam Khraisat, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, 2(1):20, 2019.
84. Egon Kidmose, Matija Stevanovic, and Jens Myrup Pedersen. Detection of malicious domains through lexical analysis. In *2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pages 1–5. IEEE, 2018.
85. Jihyun Kim, Jaehyun Kim, Huong Le Thi Thu, and Howon Kim. Long short term memory recurrent neural network classifier for intrusion detection. In *2016 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5. IEEE, 2016.
86. Levent Koc, Thomas A Mazzuchi, and Shahram Sarkani. A network intrusion detection system based on a hidden naïve bayes multiclass classifier. *Expert Systems with Applications*, 39(18):13492–13500, 2012.
87. RT Kokila, S Thamarai Selvi, and Kannan Govindarajan. Ddos detection and analysis in sdn-based environment using support vector machine classifier. In *2014 Sixth International Conference on Advanced Computing (ICoAC)*, pages 205–210. IEEE, 2014.
88. Bojan Kolosnjaji, Apostolis Zarras, George Webster, and Claudia Eckert. Deep learning for classification of malware system call sequences. In *Australasian Joint Conference on Artificial Intelligence*, pages 137–149. Springer, 2016.
89. Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, and Benjamin Turnbull. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems*, 100:779–796, 2019.
90. Manjiri V Kotpalliwar and Rakhi Wajgi. Classification of attacks using support vector machine (svm) on kddcup'99 ids database. In *2015 Fifth International Conference on Communication Systems and Network Technologies*, pages 987–990. IEEE, 2015.
91. Dennis Kügler. "man in the middle" attacks on bluetooth. In *International Conference on Financial Cryptography*, pages 149–161. Springer, 2003.
92. Rajesh Kumar, Zhang Xiaosong, Riaz Ullah Khan, Jay Kumar, and Ijaz Ahad. Effective and explainable detection of android malware based on machine learning algorithms. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pages 35–40. ACM, 2018.

93. Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):191–201, 1992.
94. Gastón L'Huillier, Alejandro Hevia, Richard Weber, and Sebastian Rios. Latent semantic analysis and keyword extraction for phishing classification. In *2010 IEEE international conference on intelligence and security informatics*, pages 129–131. IEEE, 2010.
95. Shancang Li, Li Da Xu, and Shanshan Zhao. The internet of things: a survey. *Information Systems Frontiers*, 17(2):243–259, 2015.
96. Yinhui Li, Jingbo Xia, Silan Zhang, Jiakai Yan, Xiaochuan Ai, and Kuobin Dai. An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications*, 39(1):424–430, 2012.
97. Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36(1):16–24, 2013.
98. Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. Cann: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-based systems*, 78:13–21, 2015.
99. Brian Lindauer, Joshua Glasser, Mitch Rosen, Kurt C Wallnau, and L ExactData. Generating test data for insider threat detectors. *JoWUA*, 5(2):80–94, 2014.
100. Richard P Lippmann, David J Fried, Isaac Graf, Joshua W Haines, Kristopher R Kendall, David McClung, Dan Weber, Seth E Webster, Dan Wyschogrod, Robert K Cunningham, et al. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, volume 2, pages 12–26. IEEE, 2000.
101. Hongyu Liu, Bo Lang, Ming Liu, and Hanbing Yan. Cnn and rnn based payload classification methods for attack detection. *Knowledge-Based Systems*, 163:332–341, 2019.
102. Manuel Lopez-Martin, Belen Carro, and Antonio Sanchez-Esguevillas. Application of deep reinforcement learning to intrusion detection for supervised problems. *Expert Systems with Applications*, 141:112963, 2020.
103. Rachid Ait Maalem Lahcen, Bruce Caulkins, Ram Mohapatra, and Manish Kumar. Review and insight on the behavioral aspects of cybersecurity. *Cybersecurity*, 3:1–18, 2020.
104. James MacQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 1967.
105. Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79, 2001.
106. Arif Jamal Malik and Farrukh Aslam Khan. A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection. *Cluster Computing*, 21(1):667–680, 2018.
107. Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.
108. Timothy McIntosh, Julian Jang-Jaccard, Paul Watters, and Teo Susnjak. The inadequacy of entropy-based ransomware detection. In *International Conference on Neural Information Processing*, pages 181–189. Springer, 2019.
109. Weizhi Meng, Wenjuan Li, and Lam-For Kwok. Design of intelligent knn-based alarm filter using knowledge-based alert verification in intrusion detection. *Security and Communication Networks*, 8(18):3883–3895, 2015.
110. Robert Mitchell and Ray Chen. Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems. *IEEE Transactions on Dependable and Secure Computing*, 12(1):16–30, 2014.
111. Ahmad Moghimi, Jan Wichelmann, Thomas Eisenbarth, and Berk Sunar. Memjam: A false dependency attack against constant-time crypto implementations. *International Journal of Parallel Programming*, 47(4):538–570, 2019.
112. Sara Mohammadi, Hamid Mirvaziri, Mostafa Ghazizadeh-Ahsaee, and Hadis Karimipour. Cyber intrusion detection by combined feature selection algorithm. *Journal of information security and applications*, 44:80–88, 2019.
113. Serguei A Mokhov, Joey Paquet, and Mourad Debbabi. The use of nlp techniques in static code analysis to detect weaknesses and vulnerabilities. In *Canadian Conference on Artificial Intelligence*, pages 326–332. Springer, 2014.
114. Daesung Moon, Hyungjin Im, Ikkyun Kim, and Jong Hyuk Park. Dtb-ids: an intrusion detection system based on decision tree using behavior analysis for preventing apt attacks. *The Journal of supercomputing*, 73(7):2881–2895, 2017.
115. Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 military communications and information systems conference (MilCIS)*, pages 1–6. IEEE, 2015.
116. Srinivas Mukkamala, Andrew Sung, and Ajith Abraham. Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools. *Vemuri, V. Rao, Enhancing Computer Security with Smart Technology.(Auerbach, 2006)*, pages 125–163, 2005.
117. Marc Ohm, Arnold Sykosch, and Michael Meier. Towards detection of software supply chain attacks by forensic artifacts. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, pages 1–6, 2020.
118. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
119. Teresa Pereira and Henrique Santos. An ontology based approach to information security. In *Research Conference on Metadata and Semantic Research*, pages 183–192. Springer, 2009.
120. Ian Perera, Jena Hwang, Kevin Bayas, Bonnie Dorr, and Yorick Wilks. Cyberattack prediction through public text analysis and mini-theories. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3001–3010. IEEE, 2018.
121. Muhammad Shakil Pervez and Dewan Md Farid. Feature selection and intrusion classification in nsl-kdd cup 99 dataset employing svms. In *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)*, pages 1–6. IEEE, 2014.
122. Shubha Puthran and Ketan Shah. Intrusion detection using improved decision tree algorithm with binary

and quad split. In *International Symposium on Security in Computing and Communication*, pages 427–438. Springer, 2016.
123. Hui Qi, Xiaoqiang Di, and Jinqing Li. Formal definition and analysis of access control model based on role and attribute. *Journal of information security and applications*, 43:53–60, 2018.
124. Xiaofei Qu, Lin Yang, Kai Guo, Linru Ma, Meng Sun, Mingxing Ke, and Mu Li. A survey on the development of self-organizing maps for unsupervised intrusion detection. *Mobile Networks and Applications*, pages 1–22, 2019.
125. J. Ross Quinlan. C4.5: Programs for machine learning. *Machine Learning*, 1993.
126. Daniel J Ragsdale, CA Carver, Jeffrey W Humphries, and Udo W Pooch. Adaptation techniques for intrusion detection and intrusion response systems. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, volume 4, pages 2344–2349. IEEE, 2000.
127. Kajal Rai, M Syamala Devi, and Ajay Guleria. Decision tree based algorithm for intrusion detection. *International Journal of Advanced Networking and Applications*, 7(4):2828, 2016.
128. Lee Rainie, Janna Anderson, and Jennifer Connolly. Cyber attacks likely to increase. *Digital Life in*, 2025, 2014.
129. MR Gauthama Raman, Nivethitha Somu, Sahruday Jagarapu, Tina Manghnani, Thirumaran Selvam, Kannan Krithivasan, and VS Shankar Sriram. An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm. *Artificial Intelligence Review*, pages 1–32, 2019.
130. Neha G Relan and Dharmaraj R Patil. Implementation of network intrusion detection system using variant of decision tree algorithm. In *2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE)*, pages 1–5. IEEE, 2015.
131. Lior Rokach. A survey of clustering algorithms. In *Data Mining and Knowledge Discovery Handbook*, pages 269–298. Springer, 2010.
132. Iqbal H Sarker. Context-aware rule learning from smartphone data: survey, challenges and future directions. *Journal of Big Data*, 6(1):95, 2019.
133. Iqbal H Sarker. A machine learning based robust prediction model for real-life mobile phone data. *Internet of Things*, 5:180–193, 2019.
134. Iqbal H Sarker, Yoosef B Abushark, Fawaz Alsolami, and Asif Irshad Khan. Intrudtree: A machine learning based cyber security intrusion detection model. *Symmetry*, 12(5):754, 2020.
135. Iqbal H Sarker, Yoosef B Abushark, and Asif Irshad Khan. Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. *Symmetry*, 12(4):499, 2020.
136. Iqbal H Sarker, Alan Colman, and Jun Han. Recencyminer: mining recency-based personalized behavior from contextual smartphone data. *Journal of Big Data*, 6(1):49, 2019.
137. Iqbal H Sarker, Alan Colman, Jun Han, Asif Irshad Khan, Yoosef B Abushark, and Khaled Salah. Behavdt: A behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Networks and Applications*, pages 1–11, 2019.
138. Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Individualized time-series segmentation for mining mobile phone user behavior. *The Computer Journal, Oxford University, UK*, 61(3):349–368, 2018.
139. Iqbal H Sarker, M.M. Hoque, and Kafil et al. Uddin. Mobile data science and intelligent apps: Concepts, ai-based modeling and research directions. *Mobile Networks and Applications*, pages 1–19, 2020.
140. Iqbal H Sarker and ASM Kayes. Abc-ruleminer: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications*, 168:102762, 2020.
141. Iqbal H Sarker, ASM Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, 7(1):1–29, 2020.
142. Iqbal H Sarker, ASM Kayes, and Paul Watters. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, 6(1):1–28, 2019.
143. Joshua Saxe and Hillary Sanders. Malware data science: Attack detection and attribution. 2018.
144. Harshit Saxena and Vineet Richariya. Intrusion detection in kdd99 dataset using svm-pso and feature reduction with information gain. *International Journal of Computer Applications*, 98(6), 2014.
145. Venkatesh Gauri Shankar, Mahesh Jangid, Bali Devi, and Shikha Kabra. Mobile big data: malware and its analysis. In *Proceedings of First International Conference on Smart System, Innovations and Computing*, pages 831–842. Springer, 2018.
146. Hossein Shapoorifard and Pirooz Shamsinejad. Intrusion detection using a novel hybrid method incorporating an improved knn. *Int. J. Comput. Appl.*, 173(1):5–9, 2017.
147. Aboosaleh M Sharifi, Saeed K Amirgholipour, and Alireza Pourebrahimi. Intrusion detection based on joint of k-means and knn. *Journal of Convergence Information Technology*, 10(5):42, 2015.
148. Abraham Shaw. Data breach: from notification to prevention using pci dss. *Colum. JL & Soc. Probs.*, 43:517, 2009.
149. Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, and Ali A Ghorbani. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *computers & security*, 31(3):357–374, 2012.
150. Karl Sigler. Crypto-jacking: how cyber-criminals are exploiting the crypto-currency boom. *Computer Fraud & Security*, 2018(9):12–14, 2018.
151. Peter HA Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17(1), 1957.
152. Thorvald Sorensen. method of establishing groups of equal amplitude in plant sociology based on similarity of species. *Biol. Skr.*, 5, 1948.
153. Grimm Stephan, Hitzler Pascal, and Abecker Andreas. Knowledge representation and ontologies. *Semantic Web Services: Concepts, Technologies, and Applications*, pages 51–105, 2007.
154. Nan Sun, Jun Zhang, Paul Rimba, Shang Gao, Leo Yu Zhang, and Yang Xiang. Data-driven cybersecurity incident prediction: A survey. *IEEE Communications Surveys & Tutorials*, 21(2):1744–1772, 2018.
155. Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017.
156. Arman Tajbakhsh, Mohammad Rahmati, and Abdolreza Mirzaei. Intrusion detection using fuzzy association rules. *Applied Soft Computing*, 9(2):462–469, 2009.

157. Juan E Tapiador, Agustin Orfila, Arturo Ribagorda, and Benjamin Ramos. Key-recovery attacks on kids, a keyed anomaly detection system. *IEEE Transactions on Dependable and Secure Computing*, 12(3):312–325, 2013.
158. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A Ghorbani. A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6. IEEE, 2009.
159. Mahbod Tavallaee, Natalia Stakhanova, and Ali Akbar Ghorbani. Toward credible evaluation of anomaly-based intrusion-detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(5):516–524, 2010.
160. Fei Tong and Zheng Yan. A hybrid approach of mobile malware detection in android. *Journal of Parallel and Distributed computing*, 103:22–31, 2017.
161. Lennart Van Efferen and Amr MT Ali-Eldin. A multi-layer perceptron approach for flow-based anomaly detection. In *2017 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6. IEEE, 2017.
162. Toby Velte, Anthony Velte, and Robert Elsenpeter. *Cloud computing, a practical approach*. McGraw-Hill, Inc., 2009.
163. Eduardo Viegas, Altair O Santin, Andre Franca, Ricardo Jasinski, Volnei A Pedroni, and Luiz S Oliveira. Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems. *IEEE Transactions on Computers*, 66(1):163–177, 2016.
164. Nikos Virvilis and Dimitris Gritzalis. The big four-what we did wrong in advanced persistent threat detection. In *2013 International Conference on Availability, Reliability and Security*, pages 248–254. IEEE, 2013.
165. Satyendra Vishwakarma, Vivek Sharma, and Ankita Tiwari. An intrusion detection system using knn-aco algorithm. *Int. J. Comput. Appl.*, 171(10):18–23, 2017.
166. Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.
167. Wei Wang, Ming Zhu, Xuewen Zeng, Xiaozhou Ye, and Yiqiang Sheng. Malware traffic classification using convolutional neural network for representation learning. In *2017 International Conference on Information Networking (ICOIN)*, pages 712–717. IEEE, 2017.
168. Merrill Warkentin and Robert Willison. Behavioral and policy issues in information systems security: the insider threat. *European Journal of Information Systems*, 18(2):101–105, 2009.
169. Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
170. Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. Weka: Practical machine learning tools and techniques with java implementations. 1999.
171. Yirui Wu, Dabao Wei, and Jun Feng. Network attacks detection methods based on deep learning techniques: A survey. *Security and Communication Networks*, 2020, 2020.
172. Miao Xie, Jiankun Hu, and Jill Slay. Evaluating host-based anomaly detection systems: Application of the one-class svm algorithm to adfa-ld. In *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 978–982. IEEE, 2014.
173. Miao Xie, Jiankun Hu, Xinghuo Yu, and Elizabeth Chang. Evaluating host-based anomaly detection systems: Application of the frequency-based algorithms to adfa-ld. In *International Conference on Network and System Security*, pages 542–549. Springer, 2015.
174. Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6:35365–35381, 2018.
175. Yinxing Xue, Guozhu Meng, Yang Liu, Tian Huat Tan, Hongxu Chen, Jun Sun, and Jie Zhang. Auditing anti-malware tools by evolving android malware and dynamic loading technique. *IEEE Transactions on Information Forensics and Security*, 12(7):1529–1544, 2017.
176. Manfu Yan and Zhifang Liu. A new method of transductive svm-based network intrusion detection. In *International Conference on Computer and Computing Technologies in Agriculture*, pages 87–95. Springer, 2010.
177. Chuanlong Yin, Yuefei Zhu, Jinlong Fei, and Xinzheng He. A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access*, 5:21954–21961, 2017.
178. Jun Yin. Firewall policy management, May 10 2016. US Patent 9,338,134.
179. Yali Yuan, Georgios Kaklamanos, and Dieter Hogrefe. A novel semi-supervised adaboost technique for network anomaly detection. In *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 111–114. ACM, 2016.
180. Lotfi A Zadeh. Fuzzy logic—a personal perspective. *Fuzzy sets and systems*, 281:4–20, 2015.
181. Mattia Zago, Manuel Gil Pérez, and Gregorio Martínez Pérez. Umudga: A dataset for profiling algorithmically generated domain names in botnet detection. *Data in Brief*, page 105400, 2020.
182. Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.
183. Jiong Zhang, Mohammad Zulkernine, and Anwar Haque. Random-forests-based network intrusion detection systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5):649–659, 2008.
184. Yajin Zhou and Xuxian Jiang. Dissecting android malware: Characterization and evolution. In *2012 IEEE symposium on security and privacy*, pages 95–109. IEEE, 2012.

Authors Biography

Dr. Iqbal H. Sarker received his Ph.D. under the department of Computer Science and Software Engineering from Swinburne University of Technology, Melbourne, Australia in 2018. His professional and research interests include - Data Science, Machine Learning and Deep Learning, AI, NLP, Cybersecurity Analytics, Behavioral Analytics, IoT-Smart City Technologies and Healthcare Analytics. He has published a number of peer-reviewed Journals and Conferences in top venues, such as Journals (Journal of Network and Computer

Applications – Elsevier, USA; Internet of Things – Elsevier; Journal of Big Data – Springer Nature, UK; Mobile Network and Applications – Springer, Netherlands; Sensors - Switzerland; The Computer Journal, Oxford University Press, UK; IEEE Transactions on Artificial Intelligence, and so on) and Conferences (IEEE DSAA – Canada; IEEE Percom – Greece; ACM Ubicomp – USA and Germany; ACM Mobiquitous – Australia; Springer PAKDD – Australia; Springer ADMA - China, and so on). He is a member of ACM and IEEE, and one of the research founders of the International AIQT foundation, Switzerland.