Journal of Big Data

SURVEY PAPER Open Access

Cybersecurity data science: an overview from machine learning perspective

Iqbal H. Sarker[1,2*], A. S. M. Kayes[3], Shahriar Badsha[4], Hamed Alqahtani[5], Paul Watters[3] and Alex Ng[3]

*Correspondence:
msarker@swin.edu.au
[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia
Full list of author information is available at the end of the article

Abstract

In a computing context, cybersecurity is undergoing massive shifts in technology and its operations in recent days, and data science is driving the change. Extracting *security incident patterns* or insights from cybersecurity data and building corresponding *data-driven model*, is the key to make a security system automated and intelligent. To understand and analyze the actual phenomena with data, various scientific methods, machine learning techniques, processes, and systems are used, which is commonly known as data science. In this paper, we focus and briefly discuss on *cybersecurity data science*, where the data is being gathered from relevant cybersecurity sources, and the analytics complement the *latest data-driven patterns* for providing more effective security solutions. The concept of cybersecurity data science allows making the computing process more actionable and intelligent as compared to traditional ones in the domain of cybersecurity. We then discuss and summarize a number of associated *research issues and future directions*. Furthermore, we provide a *machine learning* based *multi-layered framework* for the purpose of cybersecurity modeling. Overall, our goal is not only to discuss cybersecurity data science and relevant methods but also to focus the applicability towards data-driven intelligent decision making for protecting the systems from cyber-attacks.

Keywords: Cybersecurity, Machine learning, Data science, Decision making, Cyber-attack, Security modeling, Intrusion detection, Cyber threat intelligence

Introduction

Due to the increasing dependency on digitalization and Internet-of-Things (IoT) [1], various security incidents such as unauthorized access [2], malware attack [3], zero-day attack [4], data breach [5], denial of service (DoS) [2], social engineering or phishing [6] etc. have grown at an exponential rate in recent years. For instance, in 2010, there were less than 50 million unique malware executables known to the security community. By 2012, they were double around 100 million, and in 2019, there are more than 900 million malicious executables known to the security community, and this number is likely to grow, according to the statistics of AV-TEST institute in Germany [7]. Cybercrime and attacks can cause devastating financial losses and affect organizations and individuals as well. It's estimated that, a data breach costs 8.19 million USD for the United States and 3.9 million USD on an average [8], and the annual

cost to the global economy from cybercrime is 400 billion USD [9]. According to Juniper Research [10], the number of records breached each year to nearly triple over the next 5 years. Thus, it's essential that organizations need to adopt and implement a strong cybersecurity approach to mitigate the loss. According to [11], the national security of a country depends on the business, government, and individual citizens having access to applications and tools which are highly secure, and the capability on detecting and eliminating such cyber-threats in a timely way. Therefore, to effectively identify various cyber incidents either previously seen or unseen, and intelligently protect the relevant systems from such cyber-attacks, is a key issue to be solved urgently.

Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attack, damage, or unauthorized access [12]. In recent days, *cybersecurity* is undergoing massive shifts in technology and its operations in the context of computing, and *data science* (DS) is driving the change, where *machine learning* (ML), a core part of *"Artificial Intelligence" (AI)* can play a vital role to discover the insights from data. Machine learning can significantly change the cybersecurity landscape and data science is leading a new scientific paradigm [13, 14]. The popularity of these related technologies is increasing day-by-day, which is shown in Fig. 1, based on the data of the last five years collected from Google Trends [15]. The figure represents timestamp information in terms of a particular date in the x-axis and corresponding popularity in the range of 0 (minimum) to 100 (maximum) in the y-axis. As shown in Fig. 1, the popularity indication values of these areas are less than 30 in 2014, while they exceed 70 in 2019, i.e., more than double in terms of increased popularity. In this paper, we focus on *cybersecurity data science* (CDS), which is broadly related to these areas in terms of security data processing techniques and intelligent decision making in real-world applications. Overall, CDS is security data-focused, applies machine learning methods to quantify cyber risks, and ultimately seeks to optimize cybersecurity operations. Thus, the purpose of this paper is for those *academia and industry* people who want to study and develop a data-driven smart cybersecurity model based on machine learning techniques. Therefore, great emphasis is placed on a thorough description of various types of machine learning methods, and their relations and usage in the context of cybersecurity. This

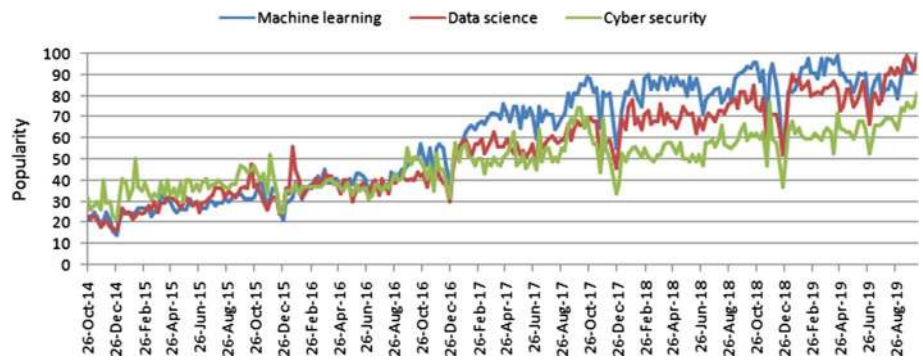

Fig. 1 Popularity trends of data science, machine learning and cybersecurity over time, where x-axis represents the timestamp information and y axis represents the corresponding popularity values

paper does not describe all of the different techniques used in cybersecurity in detail; instead, it gives an *overview* of *cybersecurity data science* modeling based on artificial intelligence, particularly from *machine learning* perspective.

The ultimate goal of cybersecurity data science is data-driven intelligent decision making from security data for smart cybersecurity solutions. CDS represents a partial paradigm shift from traditional well-known security solutions such as firewalls, user authentication and access control, cryptography systems etc. that might not be effective according to today's need in cyber industry [16–19]. The problems are these are typically handled statically by a few experienced security analysts, where data management is done in an ad-hoc manner [20, 21]. However, as an increasing number of cybersecurity incidents in different formats mentioned above continuously appear over time, such conventional solutions have encountered limitations in mitigating such cyber risks. As a result, numerous advanced attacks are created and spread very quickly throughout the Internet. Although several researchers use various data analysis and learning techniques to build cybersecurity models that are summarized in "Machine learning tasks in cybersecurity" section, a *comprehensive security model* based on the effective discovery of *security insights* and *latest security patterns* could be more useful. To address this issue, we need to develop more flexible and efficient security mechanisms that can respond to threats and to update security policies to mitigate them intelligently in a timely manner. To achieve this goal, it is inherently required to analyze a massive amount of relevant cybersecurity data generated from various sources such as network and system sources, and to discover insights or proper security policies with minimal human intervention in an automated manner.

Analyzing cybersecurity data and building the right tools and processes to successfully protect against cybersecurity incidents goes beyond a simple set of functional requirements and knowledge about risks, threats or vulnerabilities. For effectively extracting the insights or the patterns of security incidents, several machine learning techniques, such as feature engineering, data clustering, classification, and association analysis, or neural network-based deep learning techniques can be used, which are briefly discussed in "Machine learning tasks in cybersecurity" section. These learning techniques are capable to find the anomalies or malicious behavior and data-driven patterns of associated security incidents to make an intelligent decision. Thus, based on the concept of data-driven decision making, we aim to focus on *cybersecurity data science*, where the data is being gathered from relevant cybersecurity sources such as network activity, database activity, application activity, or user activity, and the analytics complement the latest data-driven patterns for providing corresponding security solutions.

The contributions of this paper are summarized as follows.

- We first make a brief discussion on the concept of *cybersecurity data science* and relevant methods to understand its applicability towards data-driven intelligent decision making in the domain of cybersecurity. For this purpose, we also make a review and brief discussion on different machine learning tasks in cybersecurity, and summarize various cybersecurity datasets highlighting their usage in different data-driven cyber applications.

- We then discuss and summarize a number of associated *research issues and future directions* in the area of cybersecurity data science, that could help both the academia and industry people to further research and development in relevant application areas.
- Finally, we provide a generic *multi-layered framework* of the cybersecurity data science model based on machine learning techniques. In this framework, we briefly discuss how the cybersecurity data science model can be used to discover useful insights from security data and making data-driven intelligent decisions to build smart cybersecurity systems.

The remainder of the paper is organized as follows. " Background" section summarizes background of our study and gives an overview of the related technologies of cybersecurity data science. " Cybersecurity data science" section defines and discusses briefly about cybersecurity data science including various categories of cyber incidents data. In " Machine learning tasks in cybersecurity" section, we briefly discuss various categories of machine learning techniques including their relations with cybersecurity tasks and summarize a number of machine learning based cybersecurity models in the field. "Research issues and future directions" section briefly discusses and highlights various research issues and future directions in the area of cybersecurity data science. In " A multi-layered framework for smart cybersecurity services" section, we suggest a machine learning-based framework to build cybersecurity data science model and discuss various layers with their roles. In " Discussion" section, we highlight several key points regarding our studies. Finally, "Conclusion" section concludes this paper.

Background

In this section, we give an overview of the related technologies of cybersecurity data science including various types of cybersecurity incidents and defense strategies.

Cybersecurity

Over the last half-century, the information and communication technology (ICT) industry has evolved greatly, which is ubiquitous and closely integrated with our modern society. Thus, protecting ICT systems and applications from cyber-attacks has been greatly concerned by the security policymakers in recent days [22]. The act of protecting ICT systems from various cyber-threats or attacks has come to be known as cybersecurity [9]. Several aspects are associated with cybersecurity: measures to protect information and communication technology; the raw data and information it contains and their processing and transmitting; associated virtual and physical elements of the systems; the degree of protection resulting from the application of those measures; and eventually the associated field of professional endeavor [23]. 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" [24]. According to Aftergood et al. [12], "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 concerns with the

understanding of diverse cyber-attacks and devising corresponding defense strategies that preserve several properties defined as below [25, 26].

- *Confidentiality* is a property used to prevent the access and disclosure of information to unauthorized individuals, entities or systems.
- *Integrity* is a property used to prevent any modification or destruction of information in an unauthorized manner.
- *Availability* is a property used to ensure timely and reliable access of information assets and systems to an authorized entity.

The term cybersecurity applies in a variety of contexts, from business to mobile computing, and can be divided into several common categories. These are - *network security* that mainly focuses on securing a computer network from cyber attackers or intruders; *application security* that takes into account keeping the software and the devices free of risks or cyber-threats; *information security* that mainly considers security and the privacy of relevant data; *operational security* that includes the processes of handling and protecting data assets. Typical cybersecurity systems are composed of network security systems and computer security systems containing a firewall, antivirus software, or an intrusion detection system [27].

Cyberattacks and security risks

The risks typically associated with any attack, which considers three security factors, such as threats, i.e., who is attacking, vulnerabilities, i.e., the weaknesses they are attacking, and impacts, i.e., what the attack does [9]. A security incident is an act that threatens the confidentiality, integrity, or availability of information assets and systems. Several types of cybersecurity incidents that may result in security risks on an organization's systems and networks or an individual [2]. These are:

- *Unauthorized access* that describes the act of accessing information to network, systems or data without authorization that results in a violation of a security policy [2];
- *Malware* known as malicious software, is any program or software that intentionally designed to cause damage to a computer, client, server, or computer network, e.g., botnets. Examples of different types of malware including computer viruses, worms, Trojan horses, adware, ransomware, spyware, malicious bots, etc. [3, 26]; Ransom malware, or *ransomware*, is an emerging form of malware that prevents users from accessing their systems or personal files, or the devices, then demands an anonymous online payment in order to restore access.
- *Denial-of-Service* is an attack meant to shut down a machine or network, making it inaccessible to its intended users by flooding the target with traffic that triggers a crash. The Denial-of-Service (DoS) attack typically uses one computer with an Internet connection, while distributed denial-of-service (DDoS) attack uses multiple computers and Internet connections to flood the targeted resource [2];
- *Phishing* a type of *social engineering*, used for a broad range of malicious activities accomplished through human interactions, in which the fraudulent attempt

takes part to obtain sensitive information such as banking and credit card details, login credentials, or personally identifiable information by disguising oneself as a trusted individual or entity via an electronic communication such as email, text, or instant message, etc. [26];

- *Zero-day attack* is considered as the term that is used to describe the threat of an unknown security vulnerability for which either the patch has not been released or the application developers were unaware [4, 28].

Beside these attacks mentioned above, privilege escalation [29], password attack [30], insider threat [31], man-in-the-middle [32], advanced persistent threat [33], SQL injection attack [34], cryptojacking attack [35], web application attack [30] etc. are well-known as security incidents in the field of cybersecurity. A *data breach* is another type of security incident, known as a data leak, which is involved in the unauthorized access of data by an individual, application, or service [5]. Thus, all data breaches are considered as security incidents, however, all the security incidents are not data breaches. Most data breaches occur in the banking industry involving the credit card numbers, personal information, followed by the healthcare sector and the public sector [36].

Cybersecurity defense strategies

Defense strategies are needed to protect data or information, information systems, and networks from cyber-attacks or intrusions. More granularly, they are responsible for preventing data breaches or security incidents and monitoring and reacting to intrusions, which can be defined as any kind of unauthorized activity that causes damage to an information system [37]. An intrusion detection system (IDS) is typically represented as "a device or software application that monitors a computer network or systems for malicious activity or policy violations" [38]. The traditional well-known security solutions such as anti-virus, firewalls, user authentication, access control, data encryption and cryptography systems, however might not be effective according to today's need in the cyber industry

[16–19]. On the other hand, IDS resolves the issues by analyzing security data from several key points in a computer network or system [39, 40]. Moreover, intrusion detection systems can be used to detect both internal and external attacks.

Intrusion detection systems are different categories according to the usage scope. For instance, a host-based intrusion detection system (HIDS), and network intrusion detection system (NIDS) are the most common types based on the scope of single computers to large networks. In a HIDS, the system monitors important files on an individual system, while it analyzes and monitors network connections for suspicious traffic in a NIDS. Similarly, based on methodologies, the signature-based IDS, and anomaly-based IDS are the most well-known variants [37].

- *Signature-based IDS*: A signature can be a predefined string, pattern, or rule that corresponds to a known attack. A particular pattern is identified as the detection of corresponding attacks in a signature-based IDS. An example of a signature can be known patterns or a byte sequence in a network traffic, or sequences used by malware. To detect the attacks, anti-virus software uses such types of sequences or pat-

terns as a signature while performing the matching operation. Signature-based IDS is also known as knowledge-based or misuse detection [41]. This technique can be efficient to process a high volume of network traffic, however, is strictly limited to the known attacks only. Thus, detecting new attacks or unseen attacks is one of the biggest challenges faced by this signature-based system.

- *Anomaly-based IDS*: The concept of anomaly-based detection overcomes the issues of signature-based IDS discussed above. In an anomaly-based intrusion detection system, the behavior of the network is first examined to find dynamic patterns, to automatically create a data-driven model, to profile the normal behavior, and thus it detects deviations in the case of any anomalies [41]. Thus, anomaly-based IDS can be treated as a dynamic approach, which follows behavior-oriented detection. The main advantage of anomaly-based IDS is the ability to identify unknown or zero-day attacks [42]. However, the issue is that the identified anomaly or abnormal behavior is not always an indicator of intrusions. It sometimes may happen because of several factors such as policy changes or offering a new service.

In addition, a hybrid detection approach [43, 44] that takes into account both the misuse and anomaly-based techniques discussed above can be used to detect intrusions. In a hybrid system, the misuse detection system is used for detecting known types of intrusions and anomaly detection system is used for novel attacks [45]. Beside these approaches, stateful protocol analysis can also be used to detect intrusions that identifies deviations of protocol state similarly to the anomaly-based method, however it uses predetermined universal profiles based on accepted definitions of benign activity [41]. In Table 1, we have summarized these common approaches highlighting their pros and cons. Once the detecting has been completed, the intrusion prevention system (IPS) that is intended to prevent malicious events, can be used to mitigate the risks in different ways such as manual, providing notification, or automatic process [46]. Among these approaches, an automatic response system could be more effective as it does not involve a human interface between the detection and response systems.

Data science

We are living in the age of data, advanced analytics, and data science, which are related to data-driven intelligent decision making. Although, the process of searching patterns or discovering hidden and interesting knowledge from data is known as data mining

Table 1 Various types of intrusion detection approaches

Approach	Pros	Cons
Signature-based IDS	Simplest and effective method to detect known attacks	Ineffective to detect unknown attacks
Anomaly-based IDS	Effective to detect new and unforeseen vulnerabilities	Anomaly is not always an indicator of intrusions, and may increase false positive rate
Hybrid approach	Reduce the false positive rate of unknown attacks	Model might be complex
Stateful protocol analysis approach	Know and trace the protocol states	Unable to inspect attacks looking like benign protocol behaviors

[47], in this paper, we use the broader term "data science" rather than data mining. The reason is that, data science, in its most fundamental form, is all about understanding of data. It involves studying, processing, and extracting valuable insights from a set of information. In addition to data mining, data analytics is also related to data science. The development of data mining, knowledge discovery, and machine learning that refers creating algorithms and program which learn on their own, together with the original data analysis and descriptive analytics from the statistical perspective, forms the general concept of "data analytics" [47]. Nowadays, many researchers use the term "data science" to describe the interdisciplinary field of data collection, preprocessing, inferring, or making decisions by analyzing the data. To understand and analyze the actual phenomena with data, various scientific methods, machine learning techniques, processes, and systems are used, which is commonly known as data science. According to Cao et al. [47] "data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments, to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology". As a high-level statement in the context of cybersecurity, we can conclude that it is the study of security data to provide data-driven solutions for the given security problems, as known as "the science of cybersecurity data". Figure 2 shows the typical data-to-insight-to-decision transfer at different periods and general analytic stages in data science, in terms of a variety of analytics goals (G) and approaches (A) to achieve the data-to-decision goal [47].

Based on the analytic power of data science including machine learning techniques, it can be a viable component of security strategies. By using data science techniques, security analysts can manipulate and analyze security data more effectively and efficiently, uncovering valuable insights from data. Thus, data science methodologies including machine learning techniques can be well utilized in the context of cybersecurity, in terms of problem understanding, gathering security data from diverse sources, preparing data to feed into the model, data-driven model building and updating, for providing smart security services, which motivates to define cybersecurity data science and to work in this research area.

Cybersecurity data science

In this section, we briefly discuss cybersecurity data science including various categories of cyber incidents data with the usage in different application areas, and the key terms and areas related to our study.

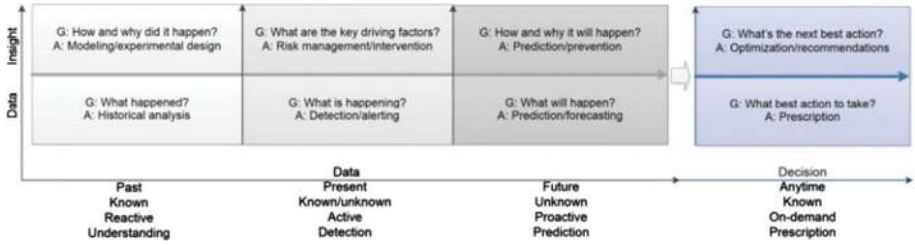

Fig. 2 Data-to-insight-to-decision analytic stages in data science [47]

Understanding cybersecurity data

Data science is largely driven by the availability of data [48]. Datasets typically represent a collection of information records that consist of several attributes or features and related facts, in which cybersecurity data science is based on. Thus, it's important to understand the nature of cybersecurity data containing various types of cyberattacks and relevant features. The reason is that raw security data collected from relevant cyber sources can be used to analyze the various patterns of security incidents or malicious behavior, to build a data-driven security model to achieve our goal. Several datasets exist in the area of cybersecurity including intrusion analysis, malware analysis, anomaly, fraud, or spam analysis that are used for various purposes. In Table 2, we summarize several such datasets including their various features and attacks that are accessible on the Internet, and highlight their usage based on machine learning techniques in different cyber applications. Effectively analyzing and processing of these security features, building target machine learning-based security model according to the requirements, and eventually, data-driven decision making, could play a role to provide intelligent cybersecurity services that are discussed briefly in "A multi-layered framework for smart cybersecurity services" section.

Defining cybersecurity data science

Data science is transforming the world's industries. It is critically important for the future of intelligent cybersecurity systems and services because of "security is all about data". When we seek to detect cyber threats, we are analyzing the security data in the form of files, logs, network packets, or other relevant sources. Traditionally, security professionals didn't use data science techniques to make detections based on these data sources. Instead, they used file hashes, custom-written rules like signatures, or manually defined heuristics [21]. Although these techniques have their own merits in several cases, it needs too much manual work to keep up with the changing cyber threat landscape. On the contrary, data science can make a massive shift in technology and its operations, where machine learning algorithms can be used to learn or extract insight of security incident patterns from the training data for their detection and prevention. For instance, to detect malware or suspicious trends, or to extract policy rules, these techniques can be used.

In recent days, the entire security industry is moving towards data science, because of its capability to transform raw data into decision making. To do this, several data-driven tasks can be associated, such as—(i) data engineering focusing practical applications of data gathering and analysis; (ii) reducing data volume that deals with filtering significant and relevant data to further analysis; (iii) discovery and detection that focuses on extracting insight or incident patterns or knowledge from data; (iv) automated models that focus on building data-driven intelligent security model; (v) targeted security alerts focusing on the generation of remarkable security alerts based on discovered knowledge that minimizes the false alerts, and (vi) resource optimization that deals with the available resources to achieve the target goals in a security system. While making data-driven decisions, behavioral analysis could also play a significant role in the domain of cybersecurity [81].

Table 2 A summary of cybersecurity datasets highlighting diverse attack-types and machine learning-based usage in different cyber applications

Dataset	Description
DARPA	Intrusion detection dataset that includes LLDOS 1.0 and LLDOS 2.0.2 attack scenario data. Data traffic and attacks containing in DARPA are collected by MIT Lincoln Laboratory for evaluating network intrusion detection systems [44, 49]
KDD'99 Cup	Most widely used data set containing 41 features for evaluating anomaly detection methods, where attacks are categorized into four major classes, such as denial of service (DoS), remote-to-local (R2L), user-to-remote (U2R), and probing [50]. KDD'99 Cup dataset can be used to evaluate ML-based attack detection model
NSL-KDD	A refined version of KDD'99 cup dataset where redundant records are eliminated. Thus ML classification based security model utilizing NSL-KDD dataset will not be biased towards more frequent records [51]
CAIDA	The datasets CAIDA'07 and CAIDA'08 contain DDoS attack traffic and normal traffic traces [52, 53]. Thus CAIDA DDoS dataset can be used to evaluate ML-based DDoS attack detection model and inferring Internet Denial-of-Service activity
ISOT'10	A combination of malicious and non-malicious type of data traffic created by Information Security and Object Technology (ISOT) research at University of Victoria [54, 55]. To evaluate ML-based classification models ISOT datasets can be used
ISCX'12	The dataset contains 19 features and 19.11% of the traffic belongs to DDoS attacks. ISCX'12 was produced at the Canadian Institute for Cybersecurity [56, 57] and can be used to evaluate the effectiveness of ML-based network intrusion detection modeling
CTU-13	A labeled malware dataset including botnet, normal, and background traffic that was captured at CTU University, Czech Republic [58]. CTU-13 can be used for data-driven malware analysis using ML techniques and to evaluate the malware detection system
UNSW-NB15	The dataset has 49 features and nine different types of attacks including DoS that was created at the University of New South Wales in 2015 [59]. UNSW-NB15 can be used for evaluating ML-based anomaly detection system in cyber applications.
CIC-IDS2018 CIC-IDS2017	The datasets include different attack scenarios, namely Brute-force, Heartbleed, Botnet, HTTP DoS, DDoS, Web attacks, and insider attack, collected by the Canadian Institute for Cybersecurity [60]. Datasets can be used for evaluating ML based intrusion detection systems including Zero-Day attacks
CIC-DDoS2019	A dataset containing DDoS attacks was collected by the Canadian Institute for Cybersecurity [61]. CIC-DDoS can be used for network traffic behavioral analytics to detect DDoS attacks using ML techniques
MAWI	A collection of Japanese network research institutions and academic institutions used to detect and evaluate DDoS intrusions using ML techniques [62]
ADFA IDS	An intrusion dataset with different versions named ADFA-LD and ADFA-WD issued by the Australian Defence Academy (ADFA) [63]. They are designed for evaluation by host-based IDS
CERT	The dataset includes users' activity logs that was created for the purpose of validating insider-threat detection systems [64, 65]. This can be used to analyze ML based user behavioral activities
Email	Email datasets are difficult to obtain because of privacy concerns. Some common corpora of emails include EnronSpam [66], SpamAssassin [67], and LingSpam [68]
DGA	The Alexa Top Sites dataset is generally used as a source of benign domain names [69]. The malicious domain names are obtained from OSINT [70] and DGArchive [71]. DGA dataset can be used for experiments in ML-based automatic DGA domains classification or botnet detection [72]
Malware	Several malware datasets such as Genome project [73], Virus Share [74], VirusTotal [75], Comodo [76], Contagio [77], DREBIN [78], and Microsoft [79] contain malicious files. These datasets can be used for data-driven malware analysis using ML techniques and to evaluate malware detection system
Bot-IoT	A dataset that incorporates legitimate and simulated IoT network traffic, along with different attacks for network forensic analytics in the area of Internet of Things [80]. Bot-IoT can be used to evaluate the reliability using different statistical and machine learning methods for forensics purposes

Thus, the concept of cybersecurity data science incorporates the methods and techniques of *data science* and *machine learning* as well as the *behavioral analytics* of various security incidents. The combination of these technologies has given birth to the term "cybersecurity data science", which refers to collect a large amount of security event data from different sources and analyze it using machine learning technologies for detecting security risks or attacks either through the discovery of useful insights or the latest data-driven patterns. It is, however, worth remembering that cybersecurity data science is not just about a collection of machine learning algorithms, rather, a process that can help security professionals or analysts to scale and automate their security activities in a smart way and in a timely manner. Therefore, the formal definition can be as follows: "Cybersecurity data science is a research or working area existing at the intersection of cybersecurity, data science, and machine learning or artificial intelligence, which is mainly security data-focused, applies machine learning methods, attempts to quantify cyber-risks or incidents, and promotes inferential techniques to analyze behavioral patterns in security data. It also focuses on

Table 3 A summary of key terms and areas related to cybersecurity data science

Key terms	Description
Security incident or attack	An incident or cyber-attack, is any act that threatens the security, confidentiality, integrity, or availability of information assets, information systems, or the networks that deliver the information
Data breach	An intentional or unintentional release of secure data to an untrusted environment, which is also known as data spill or data leak
Cyber anomaly	Anomalies are data points, items, observations or events that do not conform to the expected pattern of a given group, such as cyber intrusions or fraud. Anomalies are also referred to as outliers, noise, deviations, and exceptions in cyber data
Cybercrime	A criminal activity done using computers and the Internet, that can be committed against government and private organizations
Cybersecurity	A set of technologies and processes designed to protect networks, devices, programs, and data from various cyber attacks, damages, or unauthorized access
Data science	Focuses on the collection and application of data to provide insights or meaningful information in industry, academia, or the context of human life
Artificial intelligence (AI)	A technology that behaves intelligently with the ability of thinking and working like humans, e.g., intelligent decision making in cyber domain
Machine learning	A significant part of AI, which deals with the scientific study of algorithms and statistical models that learn from cybersecurity data to perform a specific task without using explicit instructions, relying on security incident patterns and inference instead.
Deep learning	A significant part of machine learning in AI that typically builds security models based on artificial neural networks consisting of several data processing layers
Cyber features	These are attributes, extracted from cyber data sources to analyze and build target cyber security models
Security models	Models take features as inputs and they apply simple or hybrid machine learning algorithms to come up with a specific outcome for a cybersecurity use case for intelligent decision making
Threat intelligence	Deals with gathering raw data of threats, and then analyzes and filters the data to produce usable information for automated security control systems, i.e., evidence-based knowledge in cybersecurity
Behavioral analytics	Deals with the behavioral patterns of various security incidents or the malicious behavior in the data
Internet-of-Things (IoT)	A smart environment where an object that can represent itself becomes greater by connecting to surrounding objects and the extensive data flowing around it, in which the cyber criminals are associated with.

generating security response alerts, and eventually seeks for optimizing cybersecurity solutions, to build automated and intelligent cybersecurity systems."

Table 3 highlights some key terms associated with cybersecurity data science. Overall, the outputs of cybersecurity data science are typically security data products, which can be a data-driven security model, policy rule discovery, risk or attack prediction, potential security service and recommendation, or the corresponding security system depending on the given security problem in the domain of cybersecurity. In the next section, we briefly discuss various machine learning tasks with examples within the scope of our study.

Machine learning tasks in cybersecurity

Machine learning (ML) is typically considered as a branch of "Artificial Intelligence", which is closely related to computational statistics, data mining and analytics, data science, particularly focusing on making the computers to learn from data [82, 83]. Thus, machine learning models typically comprise of a set of rules, methods, or complex "transfer functions" that can be applied to find interesting data patterns, or to recognize or predict behavior [84], which could play an important role in the area of cybersecurity. In the following, we discuss different methods that can be used to solve machine learning tasks and how they are related to cybersecurity tasks.

Supervised learning

Supervised learning is performed when specific targets are defined to reach from a certain set of inputs, i.e., task-driven approach. In the area of machine learning, the most popular supervised learning techniques are known as classification and regression methods [129]. These techniques are popular to classify or predict the future for a particular security problem. For instance, to predict denial-of-service attack (yes, no) or to identify different classes of network attacks such as scanning and spoofing, classification techniques can be used in the cybersecurity domain. ZeroR [83], OneR [130], Navies Bayes [131], Decision Tree [132, 133], K-nearest neighbors [134], support vector machines [135], adaptive boosting [136], and logistic regression [137] are the well-known classification techniques. In addition, recently Sarker et al. have proposed BehavDT [133], and IntruDtree [106] classification techniques that are able to effectively build a data-driven predictive model. On the other hand, to predict the continuous or numeric value, e.g., total phishing attacks in a certain period or predicting the network packet parameters, regression techniques are useful. Regression analyses can also be used to detect the root causes of cybercrime and other types of fraud [138]. Linear regression [82], support vector regression [135] are the popular regression techniques. The main difference between classification and regression is that the output variable in the regression is numerical or continuous, while the predicted output for classification is categorical or discrete. Ensemble learning is an extension of supervised learning while mixing different simple models, e.g., Random Forest learning [139] that generates multiple decision trees to solve a particular security task.

Unsupervised learning

In unsupervised learning problems, the main task is to find patterns, structures, or knowledge in unlabeled data, i.e., data-driven approach [140]. In the area of cybersecurity, cyber-attacks like malware stays hidden in some ways, include changing their behavior dynamically and autonomously to avoid detection. Clustering techniques, a type of unsupervised learning, can help to uncover the hidden patterns and structures from the datasets, to identify indicators of such sophisticated attacks. Similarly, in identifying anomalies, policy violations, detecting, and eliminating noisy instances in data, clustering techniques can be useful. K-means [141], K-medoids [142] are the popular partitioning clustering algorithms, and single linkage [143] or complete linkage [144] are the well-known hierarchical clustering algorithms used in various application domains. Moreover, a bottom-up clustering approach proposed by Sarker et al. [145] can also be used by taking into account the data characteristics.

Besides, feature engineering tasks like optimal feature selection or extraction related to a particular security problem could be useful for further analysis [106]. Recently, Sarker et al. [106] have proposed an approach for selecting security features according to their importance score values. Moreover, Principal component analysis, linear discriminant analysis, pearson correlation analysis, or non-negative matrix factorization are the popular dimensionality reduction techniques to solve such issues [82]. Association rule learning is another example, where machine learning based policy rules can prevent cyber-attacks. In an expert system, the rules are usually manually defined by a knowledge engineer working in collaboration with a domain expert [37, 140, 146]. Association rule learning on the contrary, is the discovery of rules or relationships among a set of available security features or attributes in a given dataset [147]. To quantify the strength of relationships, correlation analysis can be used [138]. Many association rule mining algorithms have been proposed in the area of machine learning and data mining literature, such as logic-based [148], frequent pattern based [149–151], tree-based [152], etc. Recently, Sarker et al. [153] have proposed an association rule learning approach considering non-redundant generation, that can be used to discover a set of useful security policy rules. Moreover, AIS [147], Apriori [149], Apriori-TID and Apriori-Hybrid [149], FP-Tree [152], and RARM [154], and Eclat [155] are the well-known association rule learning algorithms that are capable to solve such problems by generating a set of policy rules in the domain of cybersecurity.

Neural networks and deep learning

Deep learning is a part of machine learning in the area of artificial intelligence, which is a computational model that is inspired by the biological neural networks in the human brain [82]. Artificial Neural Network (ANN) is frequently used in deep learning and the most popular neural network algorithm is backpropagation [82]. It performs learning on a multi-layer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer. The main difference between deep learning and classical machine learning is its performance on the amount of security data increases. Typically deep learning algorithms perform well when the data volumes are large, whereas machine learning algorithms perform comparatively better on small datasets [44]. In our earlier work, Sarker et al. [129], we have illustrated the effectiveness of these approaches

considering contextual datasets. However, deep learning approaches mimic the human brain mechanism to interpret large amount of data or the complex data such as images, sounds and texts [44, 129]. In terms of feature extraction to build models, deep learning reduces the effort of designing a feature extractor for each problem than the classical machine learning techniques. Beside these characteristics, deep learning typically takes a long time to train an algorithm than a machine learning algorithm, however, the test time is exactly the opposite [44]. Thus, deep learning relies more on high-performance machines with GPUs than classical machine-learning algorithms [44, 156]. The most popular deep neural network learning models include multi-layer perceptron (MLP) [157], convolutional neural network (CNN) [158], recurrent neural network (RNN) or long-short term memory (LSTM) network [121, 158]. In recent days, researchers use these deep learning techniques for different purposes such as detecting network intrusions, malware traffic detection and classification, etc. in the domain of cybersecurity [44, 159].

Other learning techniques

Semi-supervised learning can be described as a hybridization of supervised and unsupervised techniques discussed above, as it works on both the labeled and unlabeled data. In the area of cybersecurity, it could be useful, when it requires to label data automatically without human intervention, to improve the performance of cybersecurity models. Reinforcement techniques are another type of machine learning that characterizes an agent by creating its own learning experiences through interacting directly with the environment, i.e., environment-driven approach, where the environment is typically formulated as a Markov decision process and take decision based on a reward function [160]. Monte Carlo learning, Q-learning, Deep Q Networks, are the most common reinforcement learning algorithms [161]. For instance, in a recent work [126], the authors present an approach for detecting botnet traffic or malicious cyber activities using reinforcement learning combining with neural network classifier. In another work [128], the authors discuss about the application of deep reinforcement learning to intrusion detection for supervised problems, where they received the best results for the Deep Q-Network algorithm. In the context of cybersecurity, genetic algorithms that use fitness, selection, crossover, and mutation for finding optimization, could also be used to solve a similar class of learning problems [119].

Various types of machine learning techniques discussed above can be useful in the domain of cybersecurity, to build an effective security model. In Table 4, we have summarized several machine learning techniques that are used to build various types of security models for various purposes. Although these models typically represent a learning-based security model, in this paper, we aim to focus on a comprehensive cybersecurity data science model and relevant issues, in order to build a data-driven intelligent security system. In the next section, we highlight several research issues and potential solutions in the area of cybersecurity data science.

Table 4 A summary of machine learning tasks in the domain of cybersecurity

Used Technique	Purpose	References
SVM	To classify various attacks such as DoS, Probe, U2R, and R2L	Kotpalliwar et al. [85]
SVM	Feature selection, intrusion detection and classification	Pervez et al. [86], Yan et al. [87], Li et al. [88], Raman et al. [89]
SVM	DDoS detection and analysis in SDN-based environment	Kokila et al. [90]
SVM	Evaluating host-based anomaly detection systems	Xie et al. [91]
SVM-PSO	To build intrusion detection system	Saxena et al. [92]
FCM clustering, ANN and SVM	To build network intrusion detection system	Chandrasekhar et al. [93]
KNN	Network intrusion detection system	Shapoorifard et al. [94], Vishwakarma et al. [95]
KNN	To reduce the false alarm rate	Meng et al. [96]
SVM and KNN	To build intrusion detection system	Dada et al. [97]
K-means and KNN	To build intrusion detection system	Sharifi et al. [98]
KNN and Clustering	To build intrusion detection system	Lin et al. [99]
Naive Bayes	To build an intrusion detection system for multi-class classification.	Koc et al. [100]
Decision Tree	To detect the malicious code's behavior information by running malicious code on the virtual machine and analyze the behavior information for intrusion detection.	Moon et al. [101]
Decision Tree	Feature selection and to build an effective network intrusion detection system	Ingre et al. [102], Malik et al. [103], Relan et al. [104], Rai et al. [105], Sarker et al. [106], Puthran et al. [107]
Decision Tree and KNN	Anomaly intrusion detection system	Balogun et al. [108]
Genetic Algorithm and Decision Tree	To solve the problem of small disjunct in the decision tree based intrusion detection system	Azad et al. [109]
Decision Tree and ANN	To measure the performance of intrusion detection system	Jo et al. [110]
Random Forests	To build network intrusion detection systems	Zhang et al. [111]
Association Rule	To build network intrusion detection systems	Tajbakhsh et al. [112]
Behavior Rule	To build intrusion detection system for safety critical medical cyber physical systems	Mitchell et al. [113]
Supervised	For malware detection and analysis	Alazab et al. [114], Alazab et al. [4]
Semi-supervised Adaboost	For network anomaly detection	Yuan et al. [115]
Hidden Markov Models	To build an intrusion detection system	Ariu et al. [116], Aarnes et al. [117]
Genetic Algorithm	For prevention of cyberterrorism through dynamic and evolving intrusion detection	Hansen et al. [118], Aslahi et al. [119]
Deep Learning Recurrent, RNN, LSTM	To build anomaly intrusion detection system and attack classification	Alrawashdeh et al. [120], Yin et al. [121], Kim et al. [122], Almiani et al. [123]
Deep Learning Convolutional	Malware traffic classification system	Kolosnjaji et al. [124], Wang et al. [125]
Deep and Reinforcement Learning	Malicious activities and intrusion detection system	Alauthman et al. [126], Blanco et al. [127], Lopez et al. [128]

Research issues and future directions

Our study opens several research issues and challenges in the area of cybersecurity data science to extract insight from relevant data towards data-driven intelligent decision making for cybersecurity solutions. In the following, we summarize these challenges ranging from data collection to decision making.

- *Cybersecurity datasets*: Source datasets are the primary component to work in the area of cybersecurity data science. Most of the existing datasets are old and might insufficient in terms of understanding the recent behavioral patterns of various cyber-attacks. Although the data can be transformed into a meaningful understanding level after performing several processing tasks, there is still a lack of understanding of the characteristics of recent attacks and their patterns of happening. Thus, further processing or machine learning algorithms may provide a low accuracy rate for making the target decisions. Therefore, establishing a large number of recent datasets for a particular problem domain like cyber risk prediction or intrusion detection is needed, which could be one of the major challenges in cybersecurity data science.
- *Handling quality problems in cybersecurity datasets*: The cyber datasets might be noisy, incomplete, insignificant, imbalanced, or may contain inconsistency instances related to a particular security incident. Such problems in a data set may affect the quality of the learning process and degrade the performance of the machine learning-based models [162]. To make a data-driven intelligent decision for cybersecurity solutions, such problems in data is needed to deal effectively before building the cyber models. Therefore, understanding such problems in cyber data and effectively handling such problems using existing algorithms or newly proposed algorithm for a particular problem domain like malware analysis or intrusion detection and prevention is needed, which could be another research issue in cybersecurity data science.
- *Security policy rule generation*: Security policy rules reference security zones and enable a user to allow, restrict, and track traffic on the network based on the corresponding user or user group, and service, or the application. The policy rules including the general and more specific rules are compared against the incoming traffic in sequence during the execution, and the rule that matches the traffic is applied. The policy rules used in most of the cybersecurity systems are static and generated by human expertise or ontology-based [163, 164]. Although, association rule learning techniques produce rules from data, however, there is a problem of redundancy generation [153] that makes the policy rule-set complex. Therefore, understanding such problems in policy rule generation and effectively handling such problems using existing algorithms or newly proposed algorithm for a particular problem domain like access control [165] is needed, which could be another research issue in cybersecurity data science.
- *Hybrid learning method*: Most commercial products in the cybersecurity domain contain signature-based intrusion detection techniques [41]. However, missing features or insufficient profiling can cause these techniques to miss unknown attacks. In that case, anomaly-based detection techniques or hybrid technique combining signature-based and anomaly-based can be used to overcome such issues. A hybrid technique combining multiple learning techniques or a combination of deep learning

and machine-learning methods can be used to extract the target insight for a particular problem domain like intrusion detection, malware analysis, access control, etc. and make the intelligent decision for corresponding cybersecurity solutions.

- *Protecting the valuable security information*: Another issue of a cyber data attack is the loss of extremely valuable data and information, which could be damaging for an organization. With the use of encryption or highly complex signatures, one can stop others from probing into a dataset. In such cases, cybersecurity data science can be used to build a data-driven impenetrable protocol to protect such security information. To achieve this goal, cyber analysts can develop algorithms by analyzing the history of cyberattacks to detect the most frequently targeted chunks of data. Thus, understanding such data protecting problems and designing corresponding algorithms to effectively handling these problems, could be another research issue in the area of cybersecurity data science.
- *Context-awareness in cybersecurity*: Existing cybersecurity work mainly originates from the relevant cyber data containing several low-level features. When data mining and machine learning techniques are applied to such datasets, a related pattern can be identified that describes it properly. However, a broader contextual information [140, 145, 166] like temporal, spatial, relationship among events or connections, dependency can be used to decide whether there exists a suspicious activity or not. For instance, some approaches may consider individual connections as DoS attacks, while security experts might not treat them as malicious by themselves. Thus, a significant limitation of existing cybersecurity work is the lack of using the contextual information for predicting risks or attacks. Therefore, context-aware adaptive cybersecurity solutions could be another research issue in cybersecurity data science.
- *Feature engineering in cybersecurity*: The efficiency and effectiveness of a machine learning-based security model has always been a major challenge due to the high volume of network data with a large number of traffic features. The large dimensionality of data has been addressed using several techniques such as principal component analysis (PCA) [167], singular value decomposition (SVD) [168] etc. In addition to low-level features in the datasets, the contextual relationships between suspicious activities might be relevant. Such contextual data can be stored in an ontology or taxonomy for further processing. Thus how to effectively select the optimal features or extract the significant features considering both the low-level features as well as the contextual features, for effective cybersecurity solutions could be another research issue in cybersecurity data science.
- *Remarkable security alert generation and prioritizing*: In many cases, the cybersecurity system may not be well defined and may cause a substantial number of false alarms that are unexpected in an intelligent system. For instance, an IDS deployed in a real-world network generates around nine million alerts per day [169]. A network-based intrusion detection system typically looks at the incoming traffic for matching the associated patterns to detect risks, threats or vulnerabilities and generate security alerts. However, to respond to each such alert might not be effective as it consumes relatively huge amounts of time and resources, and consequently may result in a self-inflicted DoS. To overcome this problem, a high-level management is required that correlate the security alerts considering the current context and their logical rela-

tionship including their prioritization before reporting them to users, which could be another research issue in cybersecurity data science.

- *Recency analysis in cybersecurity solutions*: Machine learning-based security models typically use a large amount of static data to generate data-driven decisions. Anomaly detection systems rely on constructing such a model considering normal behavior and anomaly, according to their patterns. However, normal behavior in a large and dynamic security system is not well defined and it may change over time, which can be considered as an incremental growing of dataset. The patterns in incremental datasets might be changed in several cases. This often results in a substantial number of false alarms known as false positives. Thus, a recent malicious behavioral pattern is more likely to be interesting and significant than older ones for predicting unknown attacks. Therefore, effectively using the concept of recency analysis [170] in cybersecurity solutions could be another issue in cybersecurity data science.

The most important work for an intelligent cybersecurity system is to develop an effective framework that supports data-driven decision making. In such a framework, we need to consider advanced data analysis based on machine learning techniques, so that the framework is capable to minimize these issues and to provide automated and intelligent security services. Thus, a well-designed security framework for cybersecurity data and the experimental evaluation is a very important direction and a big challenge as well. In the next section, we suggest and discuss a data-driven cybersecurity framework based on machine learning techniques considering multiple processing layers.

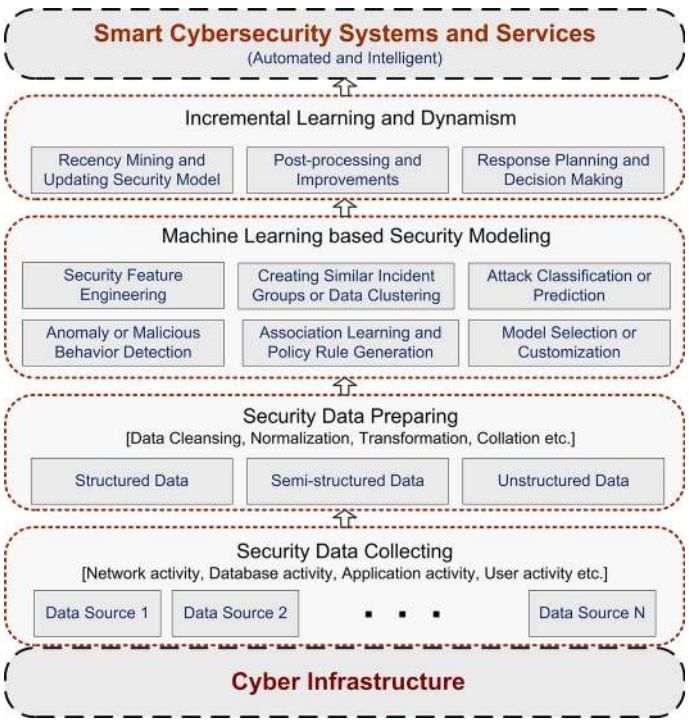

Fig. 3 A generic multi-layered framework based on machine learning techniques for smart cybersecurity services

A multi-layered framework for smart cybersecurity services

As discussed earlier, cybersecurity data science is data-focused, applies machine learning methods, attempts to quantify cyber risks, promotes inferential techniques to analyze behavioral patterns, focuses on generating security response alerts, and eventually seeks for optimizing cybersecurity operations. Hence, we briefly discuss a multiple data processing layered framework that potentially can be used to discover security insights from the raw data to build smart cybersecurity systems, e.g., dynamic policy rule-based access control or intrusion detection and prevention system. To make a data-driven intelligent decision in the resultant cybersecurity system, understanding the security problems and the nature of corresponding security data and their vast analysis is needed. For this purpose, our suggested framework not only considers the *machine learning* techniques to build the security model but also takes into account the *incremental learning and dynamism* to keep the model up-to-date and corresponding response generation, which could be more effective and intelligent for providing the expected services. Figure 3 shows an overview of the framework, involving several processing layers, from raw security event data to services. In the following, we briefly discuss the working procedure of the framework.

Security data collecting

Collecting valuable cybersecurity data is a crucial step, which forms a connecting link between security problems in cyberinfrastructure and corresponding data-driven solution steps in this framework, shown in Fig. 3. The reason is that cyber data can serve as the source for setting up ground truth of the security model that affect the model performance. The quality and quantity of cyber data decide the feasibility and effectiveness of solving the security problem according to our goal. Thus, the concern is how to collect valuable and unique needs data for building the data-driven security models.

The general step to collect and manage security data from diverse data sources is based on a particular security problem and project within the enterprise. Data sources can be classified into several broad categories such as network, host, and hybrid [171]. Within the network infrastructure, the security system can leverage different types of security data such as IDS logs, firewall logs, network traffic data, packet data, and honeypot data, etc. for providing the target security services. For instance, a given IP is considered malicious or not, could be detected by performing data analysis utilizing the data of IP addresses and their cyber activities. In the domain of cybersecurity, the network source mentioned above is considered as the primary security event source to analyze. In the host category, it collects data from an organization's host machines, where the data sources can be operating system logs, database access logs, web server logs, email logs, application logs, etc. Collecting data from both the network and host machines are considered a hybrid category. Overall, in a data collection layer the network activity, database activity, application activity, and user activity can be the possible security event sources in the context of cybersecurity data science.

Security data preparing

After collecting the raw security data from various sources according to the problem domain discussed above, this layer is responsible to prepare the raw data for building the

model by applying various necessary processes. However, not all of the collected data contributes to the model building process in the domain of cybersecurity [172]. Therefore, the useless data should be removed from the rest of the data captured by the network sniffer. Moreover, data might be noisy, have missing or corrupted values, or have attributes of widely varying types and scales. High quality of data is necessary for achieving higher accuracy in a data-driven model, which is a process of learning a function that maps an input to an output based on example input-output pairs. Thus, it might require a procedure for data cleaning, handling missing or corrupted values. Moreover, security data features or attributes can be in different types, such as continuous, discrete, or symbolic [106]. Beyond a solid understanding of these types of data and attributes and their permissible operations, its need to preprocess the data and attributes to convert into the target type. Besides, the raw data can be in different types such as structured, semi-structured, or unstructured, etc. Thus, normalization, transformation, or collation can be useful to organize the data in a structured manner. In some cases, natural language processing techniques might be useful depending on data type and characteristics, e.g., textual contents. As both the quality and quantity of data decide the feasibility of solving the security problem, effectively pre-processing and management of data and their representation can play a significant role to build an effective security model for intelligent services.

Machine learning-based security modeling

This is the core step where insights and knowledge are extracted from data through the application of cybersecurity data science. In this section, we particularly focus on machine learning-based modeling as machine learning techniques can significantly change the cybersecurity landscape. The security features or attributes and their patterns in data are of high interest to be discovered and analyzed to extract security insights. To achieve the goal, a deeper understanding of data and machine learning-based analytical models utilizing a large number of cybersecurity data can be effective. Thus, various machine learning tasks can be involved in this model building layer according to the solution perspective. These are - *security feature engineering* that mainly responsible to transform raw security data into informative features that effectively represent the underlying security problem to the data-driven models. Thus, several data-processing tasks such as feature transformation and normalization, feature selection by taking into account a subset of available security features according to their correlations or importance in modeling, or feature generation and extraction by creating new brand principal components, may be involved in this module according to the security data characteristics. For instance, the chi-squared test, analysis of variance test, correlation coefficient analysis, feature importance, as well as discriminant and principal component analysis, or singular value decomposition, etc. can be used for analyzing the significance of the security features to perform the security feature engineering tasks [82].

Another significant module is *security data clustering* that uncovers hidden patterns and structures through huge volumes of security data, to identify where the new threats exist. It typically involves the grouping of security data with similar characteristics, which can be used to solve several cybersecurity problems such as detecting anomalies, policy violations, etc. Malicious behavior or anomaly detection module is typically

responsible to identify a deviation to a known behavior, where clustering-based analysis and techniques can also be used to detect malicious behavior or anomaly detection. In the cybersecurity area, *attack classification or prediction* is treated as one of the most significant modules, which is responsible to build a prediction model to classify attacks or threats and to predict future for a particular security problem. To predict denial-of-service attack or a spam filter separating tasks from other messages, could be the relevant examples. Association learning or *policy rule generation* module can play a role to build an expert security system that comprises several IF-THEN rules that define attacks. Thus, in a problem of policy rule generation for rule-based access control system, association learning can be used as it discovers the associations or relationships among a set of available security features in a given security dataset. The popular machine learning algorithms in these categories are briefly discussed in " Machine learning tasks in cybersecurity" section. The module *model selection or customization* is responsible to choose whether it uses the existing machine learning model or needed to customize. Analyzing data and building models based on traditional machine learning or deep learning methods, could achieve acceptable results in certain cases in the domain of cybersecurity. However, in terms of effectiveness and efficiency or other performance measurements considering time complexity, generalization capacity, and most importantly the impact of the algorithm on the detection rate of a system, machine learning models are needed to customize for a specific security problem. Moreover, customizing the related techniques and data could improve the performance of the resultant security model and make it better applicable in a cybersecurity domain. The modules discussed above can work separately and combinedly depending on the target security problems.

Incremental learning and dynamism

In our framework, this layer is concerned with finalizing the resultant security model by incorporating additional intelligence according to the needs. This could be possible by further processing in several modules. For instance, the *post-processing and improvement* module in this layer could play a role to simplify the extracted knowledge according to the particular requirements by incorporating domain-specific knowledge. As the attack classification or prediction models based on machine learning techniques strongly rely on the training data, it can hardly be generalized to other datasets, which could be significant for some applications. To address such kind of limitations, this module is responsible to utilize the domain knowledge in the form of taxonomy or ontology to improve attack correlation in cybersecurity applications.

Another significant module *recency mining and updating security model* is responsible to keep the security model up-to-date for better performance by extracting the latest data-driven security patterns. The extracted knowledge discussed in the earlier layer is based on a static initial dataset considering the overall patterns in the datasets. However, such knowledge might not be guaranteed higher performance in several cases, because of incremental security data with recent patterns. In many cases, such incremental data may contain different patterns which could conflict with existing knowledge. Thus, the concept of RecencyMiner [170] on incremental security data and extracting new patterns can be more effective than the existing old patterns. The reason is that recent

security patterns and rules are more likely to be significant than older ones for predicting cyber risks or attacks. Rather than processing the whole security data again, recency-based dynamic updating according to the new patterns would be more efficient in terms of processing and outcome. This could make the resultant cybersecurity model intelligent and dynamic. Finally, *response planning and decision making* module is responsible to make decisions based on the extracted insights and take necessary actions to prevent the system from the cyber-attacks to provide automated and intelligent services. The services might be different depending on particular requirements for a given security problem.

Overall, this framework is a generic description which potentially can be used to discover useful insights from security data, to build smart cybersecurity systems, to address complex security challenges, such as intrusion detection, access control management, detecting anomalies and fraud, or denial of service attacks, etc. in the area of cybersecurity data science.

Discussion

Although several research efforts have been directed towards cybersecurity solutions, discussed in " Background" , "Cybersecurity data science", and "Machine learning tasks in cybersecurity" sections in different directions, this paper presents a comprehensive view of cybersecurity data science. For this, we have conducted a literature review to understand cybersecurity data, various defense strategies including intrusion detection techniques, different types of machine learning techniques in cybersecurity tasks. Based on our discussion on existing work, several research issues related to security datasets, data quality problems, policy rule generation, learning methods, data protection, feature engineering, security alert generation, recency analysis etc. are identified that require further research attention in the domain of cybersecurity data science.

The scope of cybersecurity data science is broad. Several data-driven tasks such as intrusion detection and prevention, access control management, security policy generation, anomaly detection, spam filtering, fraud detection and prevention, various types of malware attack detection and defense strategies, etc. can be considered as the scope of cybersecurity data science. Such tasks based categorization could be helpful for security professionals including the researchers and practitioners who are interested in the domain-specific aspects of security systems [171]. The output of cybersecurity data science can be used in many application areas such as Internet of things (IoT) security [173], network security [174], cloud security [175], mobile and web applications [26], and other relevant cyber areas. Moreover, intelligent cybersecurity solutions are important for the banking industry, the healthcare sector, or the public sector, where data breaches typically occur [36, 176]. Besides, the data-driven security solutions could also be effective in AI-based blockchain technology, where AI works with huge volumes of security event data to extract the useful insights using machine learning techniques, and block-chain as a trusted platform to store such data [177].

Although in this paper, we discuss cybersecurity data science focusing on examining raw security data to data-driven decision making for intelligent security solutions, it could also be related to big data analytics in terms of data processing and decision making. Big data deals with data sets that are too large or complex having characteristics

of high data volume, velocity, and variety. Big data analytics mainly has two parts consisting of data management involving data storage, and analytics [178]. The analytics typically describe the process of analyzing such datasets to discover patterns, unknown correlations, rules, and other useful insights [179]. Thus, several advanced data analysis techniques such as AI, data mining, machine learning could play an important role in processing big data by converting big problems to small problems [180]. To do this, the potential strategies like parallelization, divide-and-conquer, incremental learning, sampling, granular computing, feature or instance selection, can be used to make better decisions, reducing costs, or enabling more efficient processing. In such cases, the concept of cybersecurity data science, particularly machine learning-based modeling could be helpful for process automation and decision making for intelligent security solutions. Moreover, researchers could consider modified algorithms or models for handing big data on parallel computing platforms like Hadoop, Storm, etc. [181].

Based on the concept of cybersecurity data science discussed in the paper, building a data-driven security model for a particular security problem and relevant empirical evaluation to measure the effectiveness and efficiency of the model, and to asses the usability in the real-world application domain could be a future work.

Conclusion

Motivated by the growing significance of cybersecurity and data science, and machine learning technologies, in this paper, we have discussed how cybersecurity data science applies to data-driven intelligent decision making in smart cybersecurity systems and services. We also have discussed how it can impact security data, both in terms of extracting insight of security incidents and the dataset itself. We aimed to work on cybersecurity data science by discussing the state of the art concerning security incidents data and corresponding security services. We also discussed how machine learning techniques can impact in the domain of cybersecurity, and examine the security challenges that remain. In terms of existing research, much focus has been provided on traditional security solutions, with less available work in machine learning technique based security systems. For each common technique, we have discussed relevant security research. The purpose of this article is to share an overview of the conceptualization, understanding, modeling, and thinking about cybersecurity data science.

We have further identified and discussed various key issues in security analysis to showcase the signpost of future research directions in the domain of cybersecurity data science. Based on the knowledge, we have also provided a generic multi-layered framework of cybersecurity data science model based on machine learning techniques, where the data is being gathered from diverse sources, and the analytics complement the latest data-driven patterns for providing intelligent security services. The framework consists of several main phases - security data collecting, data preparation, machine learning-based security modeling, and incremental learning and dynamism for smart cybersecurity systems and services. We specifically focused on extracting insights from security data, from setting a research design with particular attention to concepts for data-driven intelligent security solutions.

Overall, this paper aimed not only to discuss cybersecurity data science and relevant methods but also to discuss the applicability towards data-driven intelligent decision making in cybersecurity systems and services from machine learning perspectives. Our analysis and discussion can have several implications both for security researchers and practitioners. For researchers, we have highlighted several issues and directions for future research. Other areas for potential research include empirical evaluation of the suggested data-driven model, and comparative analysis with other security systems. For practitioners, the multi-layered machine learning-based model can be used as a reference in designing intelligent cybersecurity systems for organizations. We believe that our study on cybersecurity data science opens a promising path and can be used as a reference guide for both academia and industry for future research and applications in the area of cybersecurity.

Abbreviations

DS: Data science; ML: Machine learning; AI: Artificial Intelligence; CDS: Cybersecurity data science; ICT: Information and communication technology; IoT: Internet of Things; DDoS: Distributed Denial of Service; IDS: Intrusion detection system; IPS: Intrusion prevention system; HIDS: Host-based intrusion detection systems; NIDS: Network Intrusion Detection Systems; SIDS: Signature-based intrusion detection system; AIDS: Anomaly-based intrusion detection system.

Acknowledgements

The authors would like to thank all the reviewers for their rigorous review and comments in several revision rounds. The reviews are detailed and helpful to improve and finalize the manuscript. The authors are highly grateful to them.

Authors' contributions

This article provides not only a discussion on cybersecurity data science and relevant methods but also to discuss the applicability towards data-driven intelligent decision making in cybersecurity systems and services. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia. [2] Chittagong University of Engineering and Technology, Chittagong, 4349, Bangladesh. [3] La Trobe University, Melbourne, VIC 3086, Australia. [4] University of Nevada, Reno, USA. [5] Macquarie University, Sydney, NSW 2109, Australia.

Received: 26 October 2019 Accepted: 21 June 2020
Published online: 01 July 2020

References

1. Li S, Da Xu L, Zhao S. The internet of things: a survey. Inform Syst Front. 2015;17(2):243–59.
2. Sun N, Zhang J, Rimba P, Gao S, Zhang LY, Xiang Y. Data-driven cybersecurity incident prediction: a survey. IEEE Commun Surv Tutor. 2018;21(2):1744–72.
3. McIntosh T, Jang-Jaccard J, Watters P, Susnjak T. The inadequacy of entropy-based ransomware detection. In: International conference on neural information processing. New York: Springer; 2019. p. 181–189
4. Alazab M, Venkatraman S, Watters P, Alazab M, et al. Zero-day malware detection based on supervised learning algorithms of api call signatures (2010)
5. Shaw A. Data breach: from notification to prevention using pci dss. Colum Soc Probs. 2009;43:517.
6. Gupta BB, Tewari A, Jain AK, Agrawal DP. Fighting against phishing attacks: state of the art and future challenges. Neural Comput Appl. 2017;28(12):3629–54.
7. Av-test institute, germany, https://www.av-test.org/en/statistics/malware/. Accessed 20 Oct 2019.
8. Ibm security report, https://www.ibm.com/security/data-breach. Accessed on 20 Oct 2019.
9. Fischer EA. Cybersecurity issues and challenges: In brief. Congressional Research Service (2014)
10. Juniper research. https://www.juniperresearch.com/. Accessed on 20 Oct 2019.
11. Papastergiou S, Mouratidis H, Kalogeraki E-M. Cyber security incident handling, warning and response system for the european critical information infrastructures (cybersane). In: International Conference on Engineering Applications of Neural Networks, p. 476–487 (2019). New York: Springer

12. Aftergood S. Cybersecurity: the cold war online. Nature. 2017;547(7661):30.
13. Hey AJ, Tansley S, Tolle KM, et al. The fourth paradigm: data-intensive scientific discovery. 2009;1:
14. Cukier K. Data, data everywhere: A special report on managing information, 2010.
15. Google trends. In: https://trends.google.com/trends/, 2019.
16. Anwar S, Mohamad Zain J, Zolkipli MF, Inayat Z, Khan S, Anthony B, Chang V. From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions. Algorithms. 2017;10(2):39.
17. Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, Karimipour H. Cyber intrusion detection by combined feature selection algorithm. J Inform Sec Appl. 2019;44:80–8.
18. Tapiador JE, Orfila A, Ribagorda A, Ramos B. Key-recovery attacks on kids, a keyed anomaly detection system. IEEE Trans Depend Sec Comput. 2013;12(3):312–25.
19. Tavallaee M, Stakhanova N, Ghorbani AA. 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)
20. Foroughi F, Luksch P. Data science methodology for cybersecurity projects. arXiv preprint arXiv:1803.04219, 2018.
21. Saxe J, Sanders H. Malware data science: Attack detection and attribution, 2018.
22. Rainie L, Anderson J, Connolly J. Cyber attacks likely to increase. Digital Life in. 2014, vol. 2025.
23. Fischer EA. Creating a national framework for cybersecurity: an analysis of issues and options. LIBRARY OF CONGRESS WASHINGTON DC CONGRESSIONAL RESEARCH SERVICE, 2005.
24. Craigen D, Diakun-Thibault N, Purse R. Defining cybersecurity. Technology Innovation. Manag Rev. 2014;4(10):13–21.
25. Council NR. et al. Toward a safer and more secure cyberspace, 2007.
26. Jang-Jaccard J, Nepal S. A survey of emerging threats in cybersecurity. J Comput Syst Sci. 2014;80(5):973–93.
27. Mukkamala S, Sung A, Abraham A. Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools. Vemuri, V. Rao, Enhancing Computer Security with Smart Technology.(Auerbach, 2006), 125–163, 2005.
28. Bilge L, Dumitraş T. 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. ACM; 2012. p. 833–44.
29. Davi L, Dmitrienko A, Sadeghi A-R, Winandy M. Privilege escalation attacks on android. In: International conference on information security. New York: Springer; 2010. p. 346–60.
30. Jovičić B, Simić D. Common web application attack types and security using asp .net. ComSIS, 2006.
31. Warkentin M, Willison R. Behavioral and policy issues in information systems security: the insider threat. Eur J Inform Syst. 2009;18(2):101–5.
32. Kügler D. "man in the middle" attacks on bluetooth. In: International Conference on Financial Cryptography. New York: Springer; 2003, p. 149–61.
33. Virvilis N, Gritzalis D. The big four-what we did wrong in advanced persistent threat detection. In: 2013 International Conference on Availability, Reliability and Security. IEEE; 2013. p. 248–54.
34. Boyd SW, Keromytis AD. Sqlrand: Preventing sql injection attacks. In: International conference on applied cryptography and network security. New York: Springer; 2004. p. 292–302.
35. Sigler K. Crypto-jacking: how cyber-criminals are exploiting the crypto-currency boom. Comput Fraud Sec. 2018;2018(9):12–4.
36. 2019 data breach investigations report, https://enterprise.verizon.com/resources/reports/dbir/. Accessed 20 Oct 2019.
37. Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity. 2019;2(1):20.
38. Johnson L. Computer incident response and forensics team management: conducting a successful incident response, 2013.
39. Brahmi I, Brahmi H, Yahia SB. A multi-agents intrusion detection system using ontology and clustering techniques. In: IFIP international conference on computer science and its applications. New York: Springer; 2015. p. 381–93.
40. Qu X, Yang L, Guo K, Ma L, Sun M, Ke M, Li M. A survey on the development of self-organizing maps for unsupervised intrusion detection. In: Mobile networks and applications. 2019;1–22.
41. Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y. Intrusion detection system: a comprehensive review. J Netw Comput Appl. 2013;36(1):16–24.
42. Alazab A, Hobbs M, Abawajy J, Alazab M. Using feature selection for intrusion detection system. In: 2012 International symposium on communications and information technologies (ISCIT). IEEE; 2012. p. 296–301.
43. Viegas E, Santin AO, Franca A, Jasinski R, Pedroni VA, Oliveira LS. Towards an energy-efficient anomaly-based intrusion detection engine for embedded systems. IEEE Trans Comput. 2016;66(1):163–77.
44. Xin Y, Kong L, Liu Z, Chen Y, Li Y, Zhu H, Gao M, Hou H, Wang C. Machine learning and deep learning methods for cybersecurity. IEEE Access. 2018;6:35365–81.
45. Dutt I, Borah S, Maitra IK, Bhowmik K, Maity A, Das S. Real-time hybrid intrusion detection system using machine learning techniques. 2018, p. 885–94.
46. Ragsdale DJ, Carver C, Humphries JW, Pooch UW. 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). IEEE; 2000. vol. 4, p. 2344–2349.
47. Cao L. Data science: challenges and directions. Commun ACM. 2017;60(8):59–68.
48. Rizk A, Elragal A. Data science: developing theoretical contributions in information systems via text analytics. J Big Data. 2020;7(1):1–26.
49. Lippmann RP, Fried DJ, Graf I, Haines JW, Kendall KR, McClung D, Weber D, Webster SE, Wyschogrod D, Cunningham RK, et al. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In: Proceedings DARPA information survivability conference and exposition. DISCEX'00. IEEE; 2000. vol. 2, p. 12–26.
50. Kdd cup 99. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. Accessed 20 Oct 2019.

51. Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the kdd cup 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications. IEEE; 2009. p. 1–6.
52. Caida ddos attack 2007 dataset. http://www.caida.org/data/ passive/ddos-20070804-dataset.xml/. Accessed 20 Oct 2019.
53. Caida anonymized internet traces 2008 dataset. https://www.caida.org/data/passive/passive-2008-dataset. Accessed 20 Oct 2019.
54. Isot botnet dataset. https://www.uvic.ca/engineering/ece/isot/ datasets/index.php/. Accessed 20 Oct 2019.
55. The honeynet project. http://www.honeynet.org/chapters/france/. Accessed 20 Oct 2019.
56. Canadian institute of cybersecurity, university of new brunswick, iscx dataset, http://www.unb.ca/cic/datasets/index.html/. Accessed 20 Oct 2019.
57. Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Comput Secur. 2012;31(3):357–74.
58. The ctu-13 dataset. https://stratosphereips.org/category/datasets-ctu13. Accessed 20 Oct 2019.
59. Moustafa N, Slay J. 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). IEEE; 2015. p. 1–6.
60. Cse-cic-ids2018 [online]. available: https://www.unb.ca/cic/ datasets/ids-2018.html/. Accessed 20 Oct 2019.
61. Cic-ddos2019 [online]. available: https://www.unb.ca/cic/datasets/ddos-2019.html/. Accessed 28 Mar 2019.
62. Jing X, Yan Z, Jiang X, Pedrycz W. Network traffic fusion and analysis against ddos flooding attacks with a novel reversible sketch. Inform Fusion. 2019;51:100–13.
63. Xie M, Hu J, Yu X, Chang E. Evaluating host-based anomaly detection systems: application of the frequency-based algorithms to adfa-ld. In: International conference on network and system security. New York: Springer; 2015. p. 542–49.
64. Lindauer B, Glasser J, Rosen M, Wallnau KC, ExactData L. Generating test data for insider threat detectors. JoWUA. 2014;5(2):80–94.
65. Glasser J, Lindauer B. Bridging the gap: A pragmatic approach to generating insider threat data. In: 2013 IEEE Security and Privacy Workshops. IEEE; 2013. p. 98–104.
66. Enronspam. https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/enron-spam/. Accessed 20 Oct 2019.
67. Spamassassin. http://www.spamassassin.org/publiccorpus/. Accessed 20 Oct 2019.
68. Lingspam. https://labs-repos.iit.demokritos.gr/skel/i-config/downloads/lingspampublic.tar.gz/. Accessed 20 Oct 2019.
69. Alexa top sites. https://aws.amazon.com/alexa-top-sites/. Accessed 20 Oct 2019.
70. Bambenek consulting—master feeds. available online:http://osint.bambenekconsulting.com/feeds/. Accessed 20 Oct 2019.
71. Dgarchive. https://dgarchive.caad.fkie.fraunhofer.de/site/. Accessed 20 Oct 2019.
72. Zago M, Pérez MG, Pérez GM. Umudga: A dataset for profiling algorithmically generated domain names in botnet detection. Data in Brief. 2020;105400.
73. Zhou Y, Jiang X. Dissecting android malware: characterization and evolution. In: 2012 IEEE Symposium on security and privacy. IEEE; 2012. p. 95–109.
74. Virusshare. http://virusshare.com/. Accessed 20 Oct 2019.
75. Virustotal. https://virustotal.com/. Accessed 20 Oct 2019.
76. Comodo. https://www.comodo.com/home/internet-security/updates/vdp/database. Accessed 20 Oct 2019.
77. Contagio. http://contagiodump.blogspot.com/. Accessed 20 Oct 2019.
78. Kumar R, Xiaosong Z, Khan RU, Kumar J, Ahad I. Effective and explainable detection of android malware based on machine learning algorithms. In: Proceedings of the 2018 international conference on computing and artificial intelligence. ACM; 2018. p. 35–40.
79. Microsoft malware classification (big 2015). arXiv:org/abs/1802.10135/. Accessed 20 Oct 2019.
80. Koroniotis N, Moustafa N, Sitnikova E, Turnbull B. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: bot-iot dataset. Future Gen Comput Syst. 2019;100:779–96.
81. McIntosh TR, Jang-Jaccard J, Watters PA. Large scale behavioral analysis of ransomware attacks. In: International conference on neural information processing. New York: Springer; 2018. p. 217–29.
82. Han J, Pei J, Kamber M. Data mining: concepts and techniques, 2011.
83. Witten IH, Frank E. Data mining: Practical machine learning tools and techniques, 2005.
84. Dua S, Du X. Data mining and machine learning in cybersecurity, 2016.
85. Kotpalliwar MV, Wajgi R. Classification of attacks using support vector machine (svm) on kddcup'99 ids database. In: 2015 Fifth international conference on communication systems and network technologies. IEEE; 2015. p. 987–90.
86. Pervez MS, Farid DM. 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). IEEE; 2014. p. 1–6.
87. Yan M, Liu Z. A new method of transductive svm-based network intrusion detection. In: International conference on computer and computing technologies in agriculture. New York: Springer; 2010. p. 87–95.
88. Li Y, Xia J, Zhang S, Yan J, Ai X, Dai K. An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Syst Appl. 2012;39(1):424–30.
89. Raman MG, Somu N, Jagarapu S, Manghnani T, Selvam T, Krithivasan K, Sriram VS. An efficient intrusion detection technique based on support vector machine and improved binary gravitational search algorithm. Artificial Intelligence Review. 2019, p. 1–32.
90. Kokila R, Selvi ST, Govindarajan K. Ddos detection and analysis in sdn-based environment using support vector machine classifier. In: 2014 Sixth international conference on advanced computing (ICoAC). IEEE; 2014. p. 205–10.

91. Xie M, Hu J, Slay J. 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). IEEE; 2014. p. 978–82.
92. Saxena H, Richariya V. Intrusion detection in kdd99 dataset using svm-pso and feature reduction with information gain. Int J Comput Appl. 2014;98:6.
93. Chandrasekhar A, Raghuveer K. 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. IEEE; 2014. p. 672–76.
94. Shapoorifard H, Shamsinejad P. Intrusion detection using a novel hybrid method incorporating an improved knn. Int J Comput Appl. 2017;173(1):5–9.
95. Vishwakarma S, Sharma V, Tiwari A. An intrusion detection system using knn-aco algorithm. Int J Comput Appl. 2017;171(10):18–23.
96. Meng W, Li W, Kwok L-F. Design of intelligent knn-based alarm filter using knowledge-based alert verification in intrusion detection. Secur Commun Netw. 2015;8(18):3883–95.
97. Dada E. A hybridized svm-knn-pdapso approach to intrusion detection system. In: Proc. Fac. Seminar Ser., 2017, p. 14–21.
98. Sharifi AM, Amirgholipour SK, Pourebrahimi A. Intrusion detection based on joint of k-means and knn. J Converg Inform Technol. 2015;10(5):42.
99. Lin W-C, Ke S-W, Tsai C-F. Cann: an intrusion detection system based on combining cluster centers and nearest neighbors. Knowl Based Syst. 2015;78:13–21.
100. Koc L, Mazzuchi TA, Sarkani S. A network intrusion detection system based on a hidden naïve bayes multiclass classifier. Exp Syst Appl. 2012;39(18):13492–500.
101. Moon D, Im H, Kim I, Park JH. Dtb-ids: an intrusion detection system based on decision tree using behavior analysis for preventing apt attacks. J Supercomput. 2017;73(7):2881–95.
102. Ingre, B., Yadav, A., Soni, A.K.: Decision tree based intrusion detection system for nsl-kdd dataset. In: International conference on information and communication technology for intelligent systems. New York: Springer; 2017. p. 207–18.
103. Malik AJ, Khan FA. A hybrid technique using binary particle swarm optimization and decision tree pruning for network intrusion detection. Cluster Comput. 2018;21(1):667–80.
104. Relan NG, Patil DR. Implementation of network intrusion detection system using variant of decision tree algorithm. In: 2015 international conference on nascent technologies in the engineering field (ICNTE). IEEE; 2015. p. 1–5.
105. Rai K, Devi MS, Guleria A. Decision tree based algorithm for intrusion detection. Int J Adv Netw Appl. 2016;7(4):2828.
106. Sarker IH, Abushark YB, Alsolami F, Khan AI. Intrudtree: a machine learning based cyber security intrusion detection model. Symmetry. 2020;12(5):754.
107. Puthran S, Shah K. Intrusion detection using improved decision tree algorithm with binary and quad split. In: International symposium on security in computing and communication. New York: Springer; 2016. p. 427–438.
108. Balogun AO, Jimoh RG. Anomaly intrusion detection using an hybrid of decision tree and k-nearest neighbor, 2015.
109. Azad C, Jha VK. Genetic algorithm to solve the problem of small disjunct in the decision tree based intrusion detection system. Int J Comput Netw Inform Secur. 2015;7(8):56.
110. Jo S, Sung H, Ahn B. A comparative study on the performance of intrusion detection using decision tree and artificial neural network models. J Korea Soc Dig Indus Inform Manag. 2015;11(4):33–45.
111. Zhan J, Zulkernine M, Haque A. Random-forests-based network intrusion detection systems. IEEE Trans Syst Man Cybern C. 2008;38(5):649–59.
112. Tajbakhsh A, Rahmati M, Mirzaei A. Intrusion detection using fuzzy association rules. Appl Soft Comput. 2009;9(2):462–9.
113. Mitchell R, Chen R. Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems. IEEE Trans Depend Secure Comput. 2014;12(1):16–30.
114. Alazab M, Venkataraman S, Watters P. Towards understanding malware behaviour by the extraction of api calls. In: 2010 second cybercrime and trustworthy computing Workshop. IEEE; 2010. p. 52–59.
115. Yuan Y, Kaklamanos G, Hogrefe D. 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. ACM; 2016. p. 111–14.
116. Ariu D, Tronci R, Giacinto G. Hmmpayl: an intrusion detection system based on hidden markov models. Comput Secur. 2011;30(4):221–41.
117. Årnes A, Valeur F, Vigna G, Kemmerer RA. Using hidden markov models to evaluate the risks of intrusions. In: International workshop on recent advances in intrusion detection. New York: Springer; 2006. p. 145–64.
118. Hansen JV, Lowry PB, Meservy RD, McDonald DM. Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection. Decis Supp Syst. 2007;43(4):1362–74.
119. Aslahi-Shahri B, Rahmani R, Chizari M, Maralani A, Eslami M, Golkar MJ, Ebrahimi A. A hybrid method consisting of ga and svm for intrusion detection system. Neural Comput Appl. 2016;27(6):1669–76.
120. Alrawashdeh K, Purdy C. Toward an online anomaly intrusion detection system based on deep learning. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). IEEE; 2016. p. 195–200.
121. Yin C, Zhu Y, Fei J, He X. A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access. 2017;5:21954–61.
122. Kim J, Kim J, Thu HLT, Kim H. Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 international conference on platform technology and service (PlatCon). IEEE; 2016. p. 1–5.
123. Almiani M, AbuGhazleh A, Al-Rahayfeh A, Atiewi S, Razaque A. Deep recurrent neural network for iot intrusion detection system. Simulation Modelling Practice and Theory. 2019;102031.

124. Kolosnjaji B, Zarras A, Webster G, Eckert C. Deep learning for classification of malware system call sequences. In: Australasian joint conference on artificial intelligence. New York: Springer; 2016. p. 137–49.
125. Wang W, Zhu M, Zeng X, Ye X, Sheng Y. Malware traffic classification using convolutional neural network for representation learning. In: 2017 international conference on information networking (ICOIN). IEEE; 2017. p. 712–17.
126. Alauthman M, Aslam N, Al-kasassbeh M, Khan S, Al-Qerem A, Choo K-KR. An efficient reinforcement learning-based botnet detection approach. J Netw Comput Appl. 2020;150:102479.
127. Blanco R, Cilla JJ, Briongos S, Malagón P, Moya JM. Applying cost-sensitive classifiers with reinforcement learning to ids. In: International conference on intelligent data engineering and automated learning. New York: Springer; 2018. p. 531–38.
128. Lopez-Martin M, Carro B, Sanchez-Esguevillas A. Application of deep reinforcement learning to intrusion detection for supervised problems. Exp Syst Appl. 2020;141:112963.
129. Sarker IH, Kayes A, Watters P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. J Big Data. 2019;6(1):1–28.
130. Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn. 1993;11(1):63–90.
131. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the eleventh conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc.; 1995. p. 338–45.
132. Quinlan JR. C4.5: Programs for machine learning. Machine Learning, 1993.
133. Sarker IH, Colman A, Han J, Khan AI, Abushark YB, Salah K. Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. Mobile Networks and Applications. 2019, p. 1–11.
134. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn. 1991;6(1):37–66.
135. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to platt's smo algorithm for svm classifier design. Neural Comput. 2001;13(3):637–49.
136. Freund Y, Schapire RE, et al: Experiments with a new boosting algorithm. In: Icml, vol. 96, p. 148–156 (1996). Citeseer
137. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J Royal Stat Soc C. 1992;41(1):191–201.
138. Watters PA, McCombie S, Layton R, Pieprzyk J. Characterising and predicting cyber attacks using the cyber attacker model profile (camp). J Money Launder Control. 2012.
139. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
140. Sarker IH. Context-aware rule learning from smartphone data: survey, challenges and future directions. J Big Data. 2019;6(1):95.
141. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Fifth Berkeley symposium on mathematical statistics and probability, vol. 1, 1967.
142. Rokach L. A survey of clustering algorithms. In: Data Mining and Knowledge Discovery Handbook. New York: Springer; 2010. p. 269–98.
143. Sneath PH. The application of computers to taxonomy. J Gen Microbiol. 1957;17:1.
144. Sorensen T. method of establishing groups of equal amplitude in plant sociology based on similarity of species. Biol Skr. 1948;5.
145. Sarker IH, Colman A, Kabir MA, Han J. Individualized time-series segmentation for mining mobile phone user behavior. Comput J. 2018;61(3):349–68.
146. Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Exp Syst Appl. 2014;41(4):1690–700.
147. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: ACM SIGMOD Record. ACM; 1993. vol. 22, p. 207–16.
148. Flach PA, Lachiche N. Confirmation-guided discovery of first-order rules with tertius. Mach Learn. 2001;42(1–2):61–95.
149. Agrawal R, Srikant R, et al: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 1994, vol. 1215, p. 487–99.
150. Houtsma M, Swami A. Set-oriented mining for association rules in relational databases. In: Proceedings of the eleventh international conference on data engineering. IEEE; 1995. p. 25–33.
151. Ma BLWHY. Integrating classification and association rule mining. In: Proceedings of the fourth international conference on knowledge discovery and data mining, 1998.
152. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: ACM Sigmod Record. ACM; 2000. vol. 29, p. 1–12.
153. Sarker IH, Salim FD. Mining user behavioral rules from smartphone data through association analysis. In: Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Melbourne, Australia. New York: Springer; 2018. p. 450–61.
154. Das A, Ng W-K, Woon Y-K. Rapid association rule mining. In: Proceedings of the tenth international conference on information and knowledge management. ACM; 2001. p. 474–81.
155. Zaki MJ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 2000;12(3):372–90.
156. Coelho IM, Coelho VN, Luz EJS, Ochi LS, Guimarães FG, Rios E. A gpu deep learning metaheuristic based model for time series forecasting. Appl Energy. 2017;201:412–8.
157. Van Efferen L, Ali-Eldin AM. A multi-layer perceptron approach for flow-based anomaly detection. In: 2017 International symposium on networks, computers and communications (ISNCC). IEEE; 2017. p. 1–6.
158. Liu H, Lang B, Liu M, Yan H. Cnn and rnn based payload classification methods for attack detection. Knowl Based Syst. 2019;163:332–41.
159. Berman DS, Buczak AL, Chavis JS, Corbett CL. A survey of deep learning methods for cyber security. Information. 2019;10(4):122.
160. Bellman R. A markovian decision process. J Math Mech. 1957;1:679–84.
161. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res. 1996;4:237–85.

162. Sarker IH. A machine learning based robust prediction model for real-life mobile phone data. Internet of Things. 2019;5:180–93.
163. Kayes ASM, Han J, Colman A. OntCAAC: an ontology-based approach to context-aware access control for software services. Comput J. 2015;58(11):3000–34.
164. Kayes ASM, Rahayu W, Dillon T. An ontology-based approach to dynamic contextual role for pervasive access control. In: AINA 2018. IEEE Computer Society, 2018.
165. Colombo P, Ferrari E. Access control technologies for big data management systems: literature review and future trends. Cybersecurity. 2019;2(1):1–13.
166. Aleroud A, Karabatis G. Contextual information fusion for intrusion detection: a survey and taxonomy. Knowl Inform Syst. 2017;52(3):563–619.
167. Sarker IH, Abushark YB, Khan AI. Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. Symmetry. 2020;12(4):499.
168. Madsen RE, Hansen LK, Winther O. Singular value decomposition and principal component analysis. Neural Netw. 2004;1:1–5.
169. Qiao L-B, Zhang B-F, Lai Z-Q, Su J-S. Mining of attack models in ids alerts from network backbone by a two-stage clustering method. In: 2012 IEEE 26th international parallel and distributed processing symposium workshops & Phd Forum. IEEE; 2012. p. 1263–9.
170. Sarker IH, Colman A, Han J. Recencyminer: mining recency-based personalized behavior from contextual smartphone data. J Big Data. 2019;6(1):49.
171. Ullah F, Babar MA. Architectural tactics for big data cybersecurity analytics systems: a review. J Syst Softw. 2019;151:81–118.
172. Zhao S, Leftwich K, Owens M, Magrone F, Schonemann J, Anderson B, Medhi D. I-can-mama: Integrated campus network monitoring and management. In: 2014 IEEE network operations and management symposium (NOMS). IEEE; 2014. p. 1–7.
173. Abomhara M, et al. Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks. J Cyber Secur Mob. 2015;4(1):65–88.
174. Helali RGM. Data mining based network intrusion detection system: A survey. In: Novel algorithms and techniques in telecommunications and networking. New York: Springer; 2010. p. 501–505.
175. Ryoo J, Rizvi S, Aiken W, Kissell J. Cloud security auditing: challenges and emerging approaches. IEEE Secur Priv. 2013;12(6):68–74.
176. Densham B. Three cyber-security strategies to mitigate the impact of a data breach. Netw Secur. 2015;2015(1):5–8.
177. Salah K, Rehman MHU, Nizamuddin N, Al-Fuqaha A. Blockchain for ai: review and open research challenges. IEEE Access. 2019;7:10127–49.
178. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inform Manag. 2015;35(2):137–44.
179. Golchha N. Big data-the information revolution. Int J Adv Res. 2015;1(12):791–4.
180. Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. 2019;6(1):44.
181. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. J Big data. 2015;2(1):21.

Publisher's Note