Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Data Mining Techniques Against Cyber Threats: A Review

Jalal Sami Issa[1], Adnan Mohsin Abdulazeez[2]
jalal.sami@auas.edu.krd, adnan.mohsin@dpu.edu.krd
[1]Department, College of Informatics University for Applied Sciences
[2]Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq

Article Information

Submitted : 19 Jun 2024
Reviewed: 22 Jun 2024
Accepted : 30 Jun 2024

Keywords

Cybersecurity, Privacy Preserving, Algorithms, Evaluation, Challenges

Abstract

Data mining is a technique used to extract useful data from existing databases. Those datasets are now being shared globally. Secure communication and confidentiality are required because data from multiple sources must be collected and stored in one central location. Data mining technology involves methods that rapidly and efficiently convert vast quantities of data into relevant insights adapted to the user's requirements. Unfortunately, the utilization of data mining expertise to acquire sensitive personal information poses a threat to individuals' privacy rights.This paper provides a review of the current techniques for preventing cyber risks and safeguarding privacy through the application of data mining. Data mining is used to examine, analyze, and figure out the structure and behavior of data mining organizations. Implementing data mining with optimal outcomes is a challenging task. In the past decade, academics have extensively studied many elements of data extraction. Therefore, it is crucial to provide published research evidence pertaining to this field. For this study, a thorough evaluation was carried out of more than thirty research papers sourced from reputable literature databases. The objective was to extract significant information regarding the field of data mining. The collected data was then used to address various study inquiries about cutting-edge extraction methodologies, data mining mechanisms in cyber dangers, data extraction procedures, algorithms, and evaluation techniques. This paper discusses various research areas and issues in data mining, serving as a valuable reference for researchers in this domain.

A. Introduction

The introduction includes background problems related to supporting theories (literature review) or previous studies (both from journals, as well as current phenomena/issues) as the basis for conducting research. The presentation of the introductory part that contains the background of the problem, theoretical basis, or related research does not have to be subtitled, but is integrated into a unified paragraph, and is presented in narrative form. At the end of the introduction, the purpose and usefulness of the research results are also explained. [Cambria 12, single space]

Data mining, also known as data extraction analysis, is a technique used to extract important information from existing databases. It serves multiple purposes in the field of security, involving national security such as surveillance, as well as cyber security tasks like virus detection. Acts of aggression targeting structures and the planned demolition of critical infrastructure, such as electrical grids and communication networks, serve as illustrations of issues associated with national security.

Cybersecurity focuses on protecting computer and network systems from malicious software, such as Trojan horses and viruses. Furthermore, data mining is employed to provide solutions such as intrusion detection and auditing. Data mining enables the rapid assessment of large datasets and the identification of concealed patterns, which is crucial for developing a potent anti-malware solution capable of detecting previously unidentified threats. The efficacy of the data mining techniques hinges on the quality of the data being utilized, as it directly influences the outcomes obtained.

Furthermore, it is highly probable that every company will face a constant influx of attackers and intruders. They focus on both large corporations and small businesses. As a result, identifying cyberassailants and assessing threat levels necessitate the use of cybersecurity defensive tools, procedures, and algorithms. Cybersecurity is a collection of algorithms and approaches used to protect the integrity of nodes, networks, and data from harm, attacks, and unauthorized access. Large enterprises possess vast amounts of data, often measured in petabytes, which includes highly valuable and sensitive information. Therefore, it is crucial to safeguard this data from unauthorized access and potential threats.

We implement cybersecurity defenses to combat three types of attacks: common attacks that exploit known vulnerabilities, advanced attacks that exploit sophisticated vulnerabilities, and developing attacks that exploit new vulnerabilities. The company has integrated the information system into multiple facets, including the production, operations, and management departments[1]. These improvements necessitate the use of dependable information security strategies and systems. Cybersecurity protections include staff knowledge, prompt intrusion detection, and analysis of unprecedented emerging danger situations within the system. Cybersecurity necessitates the mitigation of both false positive and false negative vulnerabilities.

Despite implementing different preventive measures like antivirus software, encryption, and firewalls, enterprises continue to face a significant number of intrusions. Industries such as commercial banks, credit card holders, and the telecommunications sector, among others, are particularly susceptible to the need for threat detection [2]. One of the prevalent risks is the deliberate exploitation of information technology services by attackers to advance their ideological goals. (2) The act of gaining access to confidential information without the owner's consent is another common risk. (3) The hack compromises crucial data in the fields of communication technology and medical services, significantly disrupting the e-commerce environment.

The large volume of data collection necessitates the creation of a threat intelligence system capable of recognizing patterns of criminal or anonymous activity [3]. It is critical to identify cybercrime patterns in order to effectively respond to the unexpected nature of cyberattacks and adapt to emerging trends. By promptly identifying and addressing cyber threats, organizations can mitigate cyber security risks.

The primary objective of cybersecurity is to protect computer systems and data from harmful cyber threats. Cyber threats manifest in a variety of domains and manifestations, including viruses, malware, information harvesting incidents, and application infections. The rise in cybersecurity attacks in recent years has necessitated the implementation of automated threat analysis at every level of the organization or enterprise [3]. Cybersecurity is of great concern to many enterprises because of the widespread use of data devices, which creates opportunities for cyber attackers. Due to the increasing global and organized nature of cyberattacks, the government has decided to allocate additional resources and effort towards developing strategies to mitigate various risks [2]. The financial sector plays a vital role in our nation's economy.

B. Fundamental Concepts

Cyber-attacks encompass a wider scope than the conventional concept of information operations. Information operations involve the coordinated use of electronic warfare, psychological tactics, computer networks, military deception, and security operations. These operations aim to penetrate, disrupt, destroy, or manipulate human decision-making processes. They are an important aspect of decision-making within national institutions. Figure 1 provides a detailed depiction of the structure and components of a cyber-attack. The USNM Strategy for cyberspace operations defines computer network operation as the combination of attack, defense, and utilization capabilities [4].

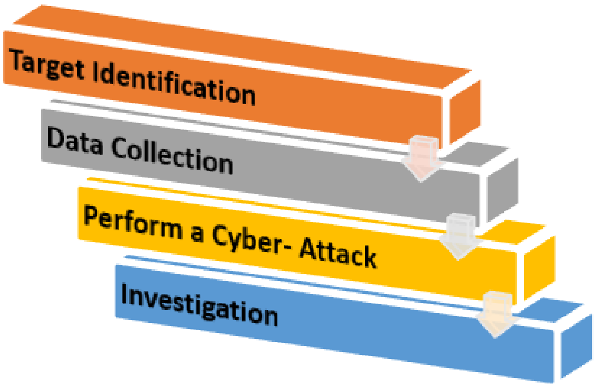

Figure 1. Analyzing of a recent cyber-attack.

This type of operation, in contrast to network attacks and network defense, primarily emphasizes the gathering and analysis of information rather than disrupting networks. Additionally, it may serve as a preliminary step in launching an attack. These activities can be conducted to distribute information and promote advertising [5]. With the increasing speed of computers and the improvement of secure failure mechanisms, it is crucial to continuously integrate cryptographic algorithms to ensure security and eliminate vulnerabilities. It is important to understand that there is a fundamental difference between cyber-crime, cyber-warfare, and cyber-attacks. As shown in Table 1. Describes the distinction between cyber-crime, cyber-warfare, and cyber-attack that defines the conceptual distinction between them.

Table 1. The differences between cyber-crime, cyber-attacks, and cyber-warfare

Type of Cyber Action	**Nature and Characteristics**
Cyber-Crime	Cyber activities exclusively carried out by non-governmental assailants. The cyber activity is executed by a computer system and solely breaches criminal law.
Cyber-Attack and Cyber-Warfare	A cyber-attack aims to sabotage and incapacitate the functioning of a computer network. The attack is likely motivated by political or security objectives.
Cyber-Warfare	The consequences of a cyber-attack are equivalent to those of an armed attack, or if the cyber incident occurs inside the framework of an armed attack.

C. Cyber Space Threats

The global cyberspace naturally encompasses several domains of control that overlap and intersect, leading to varied legal and cultural approaches and strategic goals among state actors. Nations worldwide have developed a significant reliance on cyberspace for communication and the management of tangible affairs, to the extent that it is unequivocally impracticable to detach from it. Consequently, the security responsibilities and operations of any nation are progressively influenced by the digital realm. Given the worldwide production of software and hardware items, it is unfeasible to offer assurances in the process of supplying these products. The scalability of the cyber domain sets it apart in terms of quality.

Table2. Describes the Basic Definitions and Concepts of Cyberspace [6][7].

Table 2. Describes the Basic Definitions and Concepts of Cyberspace [6][7].

Title	Definition
Cyber space	Refers to the various systems and infrastructures that are linked together, such as IT infrastructures, communication networks, computer systems, embedded processors, vital industry controllers, and information virtual environments. These networks facilitate the production, processing, storage, exchange, retrieval, and exploitation of information, and also involve the interaction between this information environment and human beings.
Cyber capital	Vital infrastructure of a country, critical cyber system, crucial information, or individuals affiliated with a country.
Cyber vulnerability	Vulnerability is a term used to describe a flaw in an asset, security processes, internal controls, or the implementation of a national cyber asset. This flaw can be taken advantage of or activated by internal or external adversaries to carry out cyber warfare.
Cyber threats	An incident capable of causing harm to missions, tasks, images, national cyber assets, or persons through an information system, by means of unauthorized access, destruction, disclosure, manipulation of information, and/or disruption of service delivery.
Cyber threat level	Cyber threats have the capability to impact several levels of infrastructure, including transnational, national, institutional, provincial, and vital cyber assets.
Probability of cyber threats	Imminent (very high), probable (high), unlikely (low), and very unlikely (very low)
Intensity of cyber threat	Levels of severity include: very high (disaster), high (crisis), moderate (major security event), low (security incident), and very low (security incident).
Cyber attack	A cyber-attack refers to any illegal action in the cyber realm that intentionally violates the security policy of a cyber-asset, resulting in harm, disruption, or interference with the services or access to information of the targeted national cyber asset. A cyber-attack refers to the deliberate utilization of a cyber-weapon to target an information system, resulting in a cyber-incident.
Cyber weapon	A cyber weapon is a purpose-built system intended to disrupt or impair the functioning or integrity of other cyber systems. These systems encompass bot networks, logic bombs, cyber vulnerability exploitation software, malware, and traffic generation systems aimed at preventing service attacks and distributed service.
Cyber warfare	Cyber warfare refers to the most advanced and intricate form of cyber-attack or cyber operation, specifically targeting a country's national cyber interests. It is characterized by its high level of sophistication and is expected to result in the most severe consequences.

Cyber warfare origin	The cyber force of the aggressor country or organizations, which are organized under the aggressor states, refers to the cyber weapons that are either controlled or abandoned by these forces.
Cyber defense	The effective utilization of all non-armed cyber and non-cyber resources within a nation to establish deterrence, prevention, rapid detection, and an effective and deterrent reaction to any cyber-attack.
Cyber biome	The term "cyber biome" refers to the creation of a natural and dynamic cyber environment that provides support for a country across multiple domains.
Virus	A virus is an autonomous program that reproduces itself and disseminates to other documents and programs through replication, perhaps resulting in program malfunctions. A computer virus functions similarly to a biological virus, disseminating itself by replicating within the cells of the host system. Notable viruses include NIMDA, SLAMMER, and SASSER.
Hacker	An individual who gains unauthorized entry into a system or enhances their level of access in order to explore, duplicate, substitute, erase, or obliterate information within it.

A limited group of individuals exercises control over several aspects of cyberspace, in addition to numerous other domains of knowledge. Users lack the capacity to alter or govern the software and gear they utilize. It is widely acknowledged that only a select few individuals has the ability to properly oversee or manipulate cyber warfare[4]. The most common vulnerabilities in the realm of cyberspace involve: external threats from foreign entities, internal threats from within the organization, threats arising from weaknesses in the supply chain of goods and services, and threats resulting from insufficient operational capacity of local forces. Foreign intelligence agencies employ cyber technologies to conduct intelligence collection and espionage operations. There have been many recorded instances worldwide of the misuse and destruction of a country's information infrastructures, which include computer systems, Internet information networks, and processors and controllers used in important industries[8]. Furthermore, there are instances where various factions, such as hackers, infiltrate the network with the intention of expressing their views. Currently, it is feasible to breach networks with minimal expertise and abilities by acquiring the requisite software and protocols from the Internet and employing them against other websites.

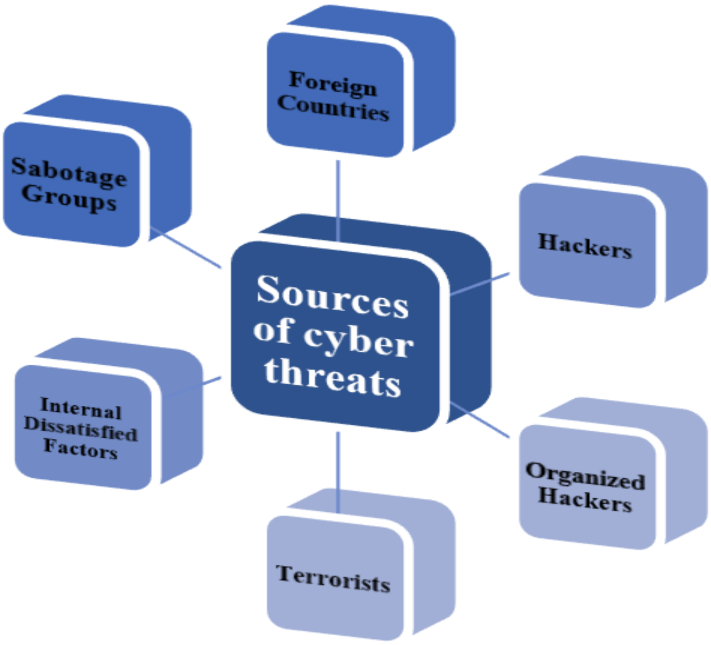

Figure 2. Shows the Sources of Cyber Threats.

The most important cyber-attack actions include Denial of Service, Logical Bomb, Abuse Tools, Sniffer, Trojan Horse, Virus, Worm, Spamming, and Botnet.

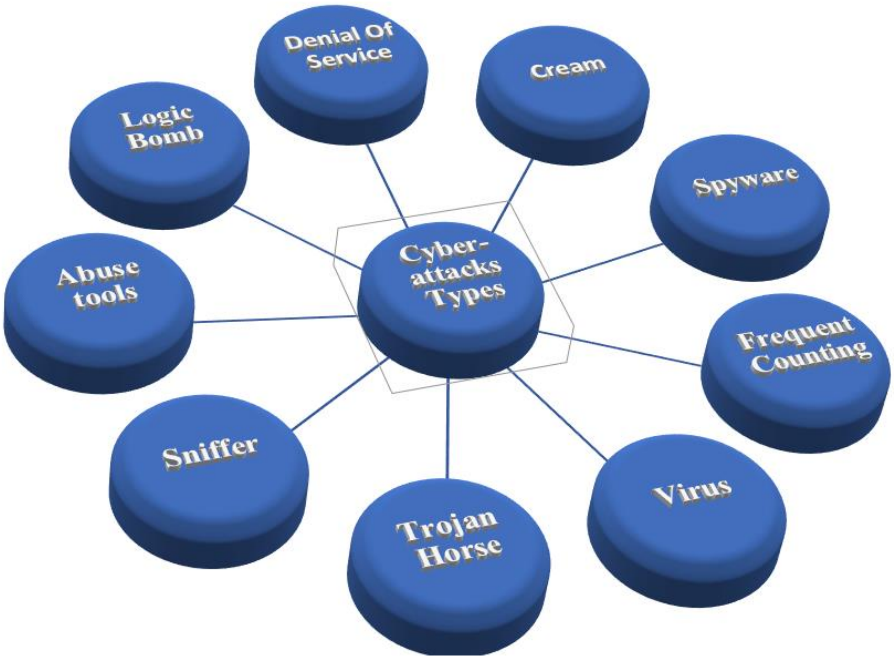

Figure 3. Types of Cyberattacks

Figure 3 illustrates the main categories of cyberattacks. The Denial of Service (DoS) method results in the loss of access for both authorized users and the system itself. Essentially, the attacker initiates a process where they inundate the target systems with many messages, consequently obstructing the lawful transmission of data. This effectively blocks any system from accessing the Internet or establishing communication with other systems[9]. In an alternative approach known as widespread Denial of Service (DoS), rather than initiating an attack from a single origin, the attackers concurrently target a substantial number of distributed systems.

D. Cyber-Security

Cybersecurity is a crucial concern in the infrastructure of all companies and organizations. Essentially, a cyber security corporation or organization can attain a prestigious position and several achievements due to its ability to safeguard sensitive and client information from rival entities. Organizations and competitors can engage in abusive behavior towards customers and individuals. First and foremost, a company or organization must prioritize providing security in an efficient way to continue to succeed and develop [10]. Cybersecurity includes practical strategies to safeguard information, networks, and data from both internal and external threats. Cybersecurity experts safeguard networks, servers, intranets, and computer systems. Cybersecurity guarantees that only persons with proper authorization can access the information. To enhance security measures, it is imperative to have a comprehensive understanding of the many categories of cyber security. Figure 5 illustrates various categories of cybersecurity.

Cybercrime refers to any illicit activity that involves illegal access or manipulation of a system, equipment, or network. There are two distinct categories of cybercrime: crimes that specifically target a system, and crimes in which a system inadvertently facilitates the criminal activity.

Table 3. Methods Commonly Used by Cybercriminals.

Method	Description	Ref.
Denial of Service	A hacker exhausts all server resources, rendering the service inaccessible to system users.	C. Topping (2021)
Man-in-the-Middle	A hacker engages in a technique known as "man-in-the-middle" attack, when they position themselves between the victim's device and the router in order to intercept or modify data packets.	Huang et al. (2020)
Malware	Malware is a method by which individuals encounter worms or viruses, resulting in the infection of their gadgets.	Manz (2020)
Phishing	Phishing is a technique employed by hackers when they send an email that appears to be genuine, but is actually a means to trick victims into revealing sensitive information.	Gayathri and Saxena (2021)

Table 3 displays the frequently employed techniques utilized by cybercriminals. Confidentiality, integrity, and availability are the fundamental concepts that form the basis of security in any company. The idea of secrecy stipulates that sensitive information and functions can only be accessed by authorized parties. Example: Classified military information (Confidentiality)[11]. The principles of integrity stipulate that only individuals and resources with proper authorization are permitted to edit, add, or remove sensitive information and functions.

E. Data Mining Mechanism in Cybersecurity Solutions

Data mining is the process of posing queries to large quantities of data and extracting valuable information, frequently of an undisclosed nature, by the application of mathematical, statistical, and machine learning methodologies. Data mining is widely used in several fields such as marketing and sales, web and e-commerce, medicine, law, manufacturing, and, more lately, national and cyber security. By employing data mining techniques, it is possible to reveal concealed interconnections among terrorist organizations and maybe anticipate future terrorist incidents by analyzing historical data. Moreover, data mining techniques can be utilized to enhance e-commerce by targeting certain audiences. Data mining can be utilized in the realm of multimedia, encompassing tasks such as video analysis and image categorization.

Data mining has practical implications in security, specifically in identifying suspicious events and detecting malicious software. This section specifically examines data mining technologies used for applications in intrusion detection, picture categorization, and online browsing. Furthermore, concentrate solely on the data mining tools that we have created specifically for cyber security purposes.

Data mining employs algorithms and diverse methodologies to transform extensive datasets into valuable output. Common data mining techniques encompass association rules, classification, clustering, decision trees, K-Nearest Neighbor, neural networks, and predictive analysis.

Data mining facilitates researchers in analyzing information, uncovering novel patterns and data, and forecasting future trends, hence enhancing the identification of malware, network breaches, insider attacks, and various other security risks. Data mining techniques have experienced a significant growth in the last ten years, and they currently offer a wide range of tools and products for various applications[2].

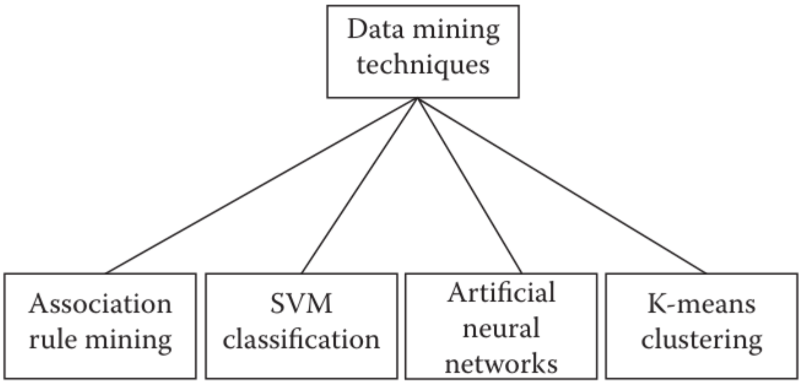

Figure 5. Data Mining Techniques.

Figure 5. illustrates the data mining strategies. Data mining has been extensively employed in various domains such as healthcare, e-commerce, and security.

Table 4. Data Mining Algorithms as Solutions in Cyber Security Problems and Its Application Domain

Ref.	Based Model	Application Domain	Cybersecurity Problems Solutions
Khan et al, 2020 [12]	Naïve Bayes, Random Forest and Sequential Minimal	Optimization ransomware family.	Detect ransomware accurately and effectively.
Gu and Lu, 2021 [13]	Naïve Bayes	Effective learning algorithms.	Intrusion detection system
Liu et al, 2022 [14]	kNN	WSN intelligent intrusion detection	Intrusion detection system
Wen-Tao Hao et al, 2022 [15]	C5.0	Data type imbalance	Intrusion detection system
Islam et al, 2022 [16]	Deep neural networks	Bio-cyber-attacks.	Trojan attacks detection
Farrukh Arslan et al, 2023 [17]	Deep Learning	Anomaly detection in time series.	Trojan attacks detection suggest future directions
Farbiash &Rami 2020, [18]	Apply machine learning/deep learning	Malware obfuscation techniques.	Detection of malicious DNA injection.
Snigdha et al, 2020 [19]	Support vector machine and K Nearest Algorithm	Biomedical	Recognized different OSA events from ERCS & breathing rate measurement
Tawalbeh et al., 2021 [20]	Deep neural network	Agriculture	Safe farm environments
AL MOGBIL et al., 2020 [21]	Machine learning	Automation	Home/ Smart city safeguard

Malware, or malicious software, is created by hackers with the intention of stealing data and identity, causing harm to systems, and denying legitimate services to users, among other destructive activities. Malware has been a persistent problem in society and the software industry for nearly forty years. Malware encompasses a range of malicious software, including as viruses, worms, Trojan horses, time and logic bombs, botnets, and spyware, various types of malwares are illustrated in Figure 6.

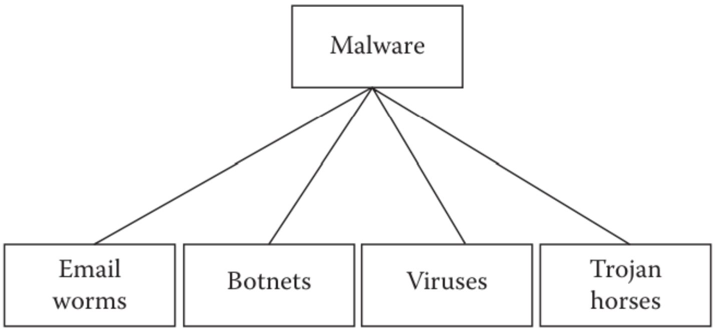

Figure 6. Illustrate Various Types of Malwares

An email worm propagates by spreading through infected email messages. The worm can be sent as an attachment, or the email may include hyperlinks to a website that is infected. Upon the user's action of opening the attachment or clicking the link, the host becomes infected instantaneously. The worm capitalizes on the susceptible email software on the host PC to dispatch infected emails to addresses stored in the address book. Therefore, newly acquired machines become contaminated. Worms inflict harm onto systems and individuals through a variety of means.

Malicious code poses a significant danger to both individual computers and the broader computer civilization. There are various types of malicious software that exist in the wild. A widely used method employed by the antivirus industry to identify dangerous code is known as "signature detection." This technique involves comparing the executables with a distinctive text or byte pattern known as a signature. The signature serves as an identifier for a certain malicious code.

Identification Botnets pose a significant danger due to their large scale and formidable capabilities. A network of infected hosts, known as botnets, is managed by a human operator called a bot-master or bot-herder from a central command and control center. The bot-master has the ability to command these bots to enlist additional bots, initiate synchronized distributed denial of service (DDoS) assaults on targeted hosts, pilfer confidential data from compromised workstations, dispatch large volumes of unsolicited spam emails, and do similar actions.

F. Data Mining Classification

Classification techniques include decision trees, support vector machines, and memory-based reasoning. Association rule mining techniques are commonly employed to establish associations. Link analysis, which examines the connections between links, can also establish correlations between linkages and anticipate the formation of new relationships.

Various techniques can be employed in clustering, each possessing distinct characteristics and being appropriate for specific situations. Some of the aspects that clustering algorithms consider include the distance between data points, the density of areas, and the distribution of data [22]. Clustering is an automated procedure; however, it may require modifications in preprocessing on occasion.

I. K-Means Clustering Algorithm

This algorithm can be described as the process of analyzing clusters, where a certain number of observations are separated into K clusters. Therefore, each observation is associated with the cluster that has the closest mean. Verona cells were created by the process of dividing the data space. K-means is widely regarded as one of the simplest machine learning algorithms that offers a solution to the widely recognized clustering problem[23].

II. EM Algorithm for Privacy

An Expectation-Maximization (EM) algorithm is a redundant technique used to compute maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models that involve unobserved hidden variables. The EM algorithm is employed once the K-means algorithm produces a good result. The Expectation Maximization algorithm involves alternating between an E-step, which calculates the expected value of the logarithm using the most recent parameter estimate, and a Maximization step. Each cluster's probability is determined by the probability distribution supplied by the EM Algorithm.

III. Hierarchical Clustering

Clusters are formed in a hierarchical manner and can be analyzed in a sequential manner. Hierarchical clustering is conducted in two distinct categories: Divisive and agglomerative. Agglomerative clustering, often known as the bottom-up approach, is a method where data points are gradually merged together. This approach terminates when the desired number of clusters, k, is obtained by the combining of several clusters. The divisive strategy involves starting clustering from one end and iteratively separating clusters one by one in a hierarchical manner. Agglomerative clustering can cause large data sets to have decreased speed.

G. Motivation and Related Work

The evaluation of cyber threat mechanisms is a significant area of research within the field of data mining. Data mining research, which began in the 2000s, has gained significant pace and attention during the past decade. The main factor is the significant growth in the size of application software, which can be attributed to the inclusion of additional functionalities and features. Furthermore, the progress of artificial intelligence (AI) techniques, including data mining, clustering, optimization, and machine learning approaches, over the past ten years has significantly contributed to the growth of research in software module clustering[24]. Various academics have developed diverse models tailored to specific sorts of datasets, which are compatible with a range of machine learning methods. Therefore, each distinct cluster exhibits substantial variations in their features.

Stefano Silvestri et al [25] Proposed machine learning for NLP-based healthcare ecosystem cyber threat and vulnerability analysis. Machine Learning algorithms like the BERT neural language model and XGBoost pull updated information from natural language documents readily available on the web to assess threats and vulnerabilities. Alberto Mozo et al [26] Describes how machine learning is used to safeguard cloud-hosted SDN controller Tera-Flow SDN from cyberattacks. The study integrates machine learning components in a distributed setting to provide secure end-to-end cyber threat defense. This demonstration used crypto mining malware, an emerging attack vector. Mohamed Abushwereb et al [27] Uses SNMP-MIB characteristics to test machine learning classification techniques for DoS detection. The study examines the most common DoS assaults and their detection probability using classifiers. Results show that SNMP-MIB-based machine learning categorization can accurately detect most DoS assaults. Hari Krishna and D. K. Sattivada [28] Proposed a bi-objective optimization problem to build accurate and exact SVMs to detect cyberattacks in a data-rich setting. An ISCX dataset intrusion detection case study used the proposed approach. The paper also addressed SVMs, their features, and optimization issues. Moreover, Serhii Toliupa et al [29] Proposed a data mining and AI approach to detect and prevent network intrusions. The methods support vector machines, fuzzy logic, clustering, and neural networks. Experimental methods included using the KDD database for intrusion detection and tiered structures for distributed information system cyberattack recognition. The results show that the suggested model for securing information networks detects network threats effectively. Ratnesh Kumar Dubey et al [30] Discuss how machine learning is used in computer defense to detect malware, fraud, and intrusion. The authors recommend selecting protective function parameters, feature discovery techniques, and classifiers. They also examine AI and machine learning's cyber security effects. The findings examine how machine learning might improve cyber security threat identification and response. Alfredo Cuzzocrea et al [31] Novel DDoS detection method employing non-linear IP address sequence analysis. Unexpected requests were detected by banning IP addresses other than those predicted by the Probability Density Function of IP address sequences. B. Muruge shwari et al [32] Propose a precise elliptic curve cryptography (PECC)-based data mining method to protect data privacy and improve algorithm performance. Experimental results show that the proposed method maintains data integrity and security while being practical and applicable in actual events. Nancy Mohamed and Magdy M. A. Salama [33] Provides a data mining-based cyber-physical assault detection solution for smart grid overcurrent relays to resist cyber attackers' compromised relay settings. Testing shows that the tool efficiently classifies settings using physical attributes exclusively. Grid cyber-physical resilience and power network security from adaptive relay cyber-physical attacks are improved by the suggested strategy. Mostofa Ahsan et al [34] Dynamic Feature Selector (DFS) was proposed and implemented to filter inconsequential variables for machine learning-based intrusion detection systems. Four machine-learning methods were used to test DFS on two meta-learning datasets. On the NSL-KDD dataset, accuracy improved from 99.54% to 99.64%, and one-hot encoded features decreased from 123 to 50. Muhammad Shoaib Akhtar and Tao Feng [35] As malware evolves, dynamic malware detection is becoming more important. This study compares machine learning techniques for this purpose. The researchers found that the RF, SGD, extra trees, and Gaussian NB classifiers had 100% accuracy, perfect precision, and good recall after analyzing test and experimental data. Steven Jorgensen et al [36] Presented a machine-learning framework for encrypted network traffic classification and contextualized uncertainty quantification for predictions. The approach was very accurate on a fresh public dataset of labeled VPN-encrypted network traffic from ten applications across five application categories. This work fills a traffic classification machine learning application gap. Majedkan, N. A. et al [37] Presented model focuses on fast diagnosis waiting times and medical service patient satisfaction in the region. A queuing system has two components: its population and its service system. This enhances will examine queuing line methods, outbreak prevention in outlets and gathering settings, and performance evaluation. Haval T. Sadeeq and Adnan M. Abdulazeez[38] analysis review of more than one hundred metaheuristics have been made. This article provides important metaheuristic insights, proposes the overall mathematical framework of MH algorithms, and divides it into tasks with probable advancement. In recent years, this profession has made great progress, yet many questions remain. Thus, new methods are proposed to address these issues.

Table 4. Representative latest works on data mining techniques and contribution

Ref	**Year**	**Based Model**	**Key Contribution**
[25]	2023	Machine Learning models (BERT neural language model and XG-Boost)	To contributes towards an effective threats and vulnerabilities
[26]	2023	Machine-Learning components in a distributed scenario	To Secure end-to-end protection against cyber threats
[27]	2020	Machine learning classification algorithms	To detected DoS attacks with high accuracy
[28]	2023	Support vector machines	To detect various cyber-attacks in a context with a lot of data
[29]	2023	Data mining methods and artificial intelligence	To protecting information in networks
[30]	2023	Machine Learning Techniques	To detection cyber attack
[31]	2021	Non-linear analysis of IP address	To innovative approach for detecting Distributed Denial of Service (DDoS) attacks
[32]	2023	Elliptic curve cryptography	To preserve data privacy while improving mining algorithm performance
[33]	2022	Cyber-physical attack detection tool	To enhance grid cyber-physical resilience and protect power networks from moving into insecure states
[34]	2021	Dynamic feature selector	To filter insignificant variables for machine learning-based intrusion detection systems
[35]	2023	Machine learning algorithms	To detection dynamic malware, which is becoming

			more crucial and sophisticated
[36]	2024	machine-learning framework	To classifying encrypted network traffic and providing contextualized uncertainty quantification for predictions.

H. Threats to Validity

In this literature review, a technique for data mining against cyber threats is discussed as follows: Data mining serves as a defence against cyberattacks. Considering that we are using different scholar databases as sources, the search strategy that we are using to discover the relevant publications might not be exhaustive. On the other hand, there may be other studies in other scholarly databases. The search process will be completed by the year 2020. The papers contain articles that were published after the date in question. It is possible to overlook research that is pertinent to the topic at hand because the process of identifying and applying inclusion and exclusion criteria is subjective. This review paper draws conclusions based on the 36 papers that were selected. Every literature mapping study has a number of threats that might affect its validity.

In this review, we eliminated many risks based on well-known thoughts, and we set the following criteria for literature illustrations:

Relevant studies covered: Data mining studies may not be entirely identifiable. Some unidentified documents may remain. To avoid missing important linked papers, the gathering technique was heavily used.

Relevant studies are covered: Furthermore, all data mining studies are identifiable. To avoid missing important related studies, the gathering strategy was heavily used.

Article exclusion/inclusion criteria: Limited author assessment and judgment can affect the application of the criteria. This study included or eliminated papers only after the authors reached an agreement to permanently resolve this problem.

I. Conclusion

Cyberspace and related technologies are one of the most important sources of power in the third millennium. As expected, this will not threaten government security. This effect can be measured in a variety of ways. This effect can be evaluated in several ways. The concept of security comes first. Today, national security is threatened by citizens' diminishing quality of life, not military challenges or internal and external frontiers. Second, cyber threats no longer have a geographical component. Third, cyberthreat vulnerability. Cyberthreat vulnerabilities, which are sporadic, complex, and often linked to important networks and infrastructure, have the potential to cause significant damage. Fourth, as previously stated, cyber risks affect both governments, individuals, and companies. Nearly every cybersecurity solution relies on trustworthy, relevant, and well-structured data. Organizations generate massive amounts of data daily, but manually gathering, analyzing, and understanding it to address cybersecurity issues is impossible. Information mining for intrusions can reveal confidential information. Clustering strategies can protect cyber security against harmful malware. With the EM algorithm, privacy is maintained without affecting computation or communication accuracy. EM usually works with real-world data.

J. References

[1] M. R. Bastos and J. S. C. Martini, "A model-free voltage stability security assessment method using artificial intelligence," in 2015 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT LATAM), 2015, pp. 679–682. doi: 10.1109/ISGT-LA.2015.7381238.

[2] F. Q. Kareem, "INTEGRATION OF CLOUD AND PARALLEL COMPUTATIONS WITH EFFICIENCY OF DATA MINING AND INTERNET OF THINGS BASED ON PRINCIPLES OF WEB TECHNOLOGY AND DISTRIBUTED SYSTEMS," pp. 1639–1689, doi: 10.17605/OSF.IO/3QRVX.

[3] D. Sisiaridis and O. Markowitch, "Reducing data complexity in feature extraction and feature selection for big data security analytics," in 2018 1st International Conference on Data Intelligence and Security (ICDIS), IEEE, 2018, pp. 43–48.

[4] L. Ma, Y. Zhang, C. Yang, and L. Zhou, "Security control for two-time-scale cyber physical systems with multiple transmission channels under DoS attacks: The input-to-state stability," J. Franklin Inst., vol. 358, no. 12, pp. 6309–6325, 2021, doi: 10.1016/j.jfranklin.2021.05.017.

[5] Y. Cao, Z. Huang, C. Ke, J. Xie, and J. Wang, "A topology-aware access control model for collaborative cyber-physical spaces: Specification and verification," Comput. Secur., vol. 87, p. 101478, 2019, doi: 10.1016/j.cose.2019.02.013.

[6] J. Ashraf et al., "IoTBoT-IDS: A novel statistical learning-enabled botnet detection framework for protecting networks of smart cities," Sustain. Cities Soc., vol. 72, no. April, p. 103041, 2021, doi: 10.1016/j.scs.2021.103041.

[7] A. Ahmed Jamal, A. A. Mustafa Majid, A. Konev, T. Kosachenko, and A. Shelupanov, "A review on security analysis of cyber physical systems using Machine learning," Mater. Today Proc., vol. 80, no. xxxx, pp. 2302–2306, 2023, doi: 10.1016/j.matpr.2021.06.320.

[8] M. Beechey, K. G. Kyriakopoulos, and S. Lambotharan, "Evidential classification and feature selection for cyber-threat hunting," Knowledge-Based Syst., vol. 226, p. 107120, 2021, doi: 10.1016/j.knosys.2021.107120.

[9] C. Topping, A. Dwyer, O. Michalec, B. Craggs, and A. Rashid, "Beware suppliers bearing gifts!: Analysing coverage of supply chain cyber security in critical national infrastructure sectorial and cross-sectorial frameworks," Comput. Secur., vol. 108, p. 102324, 2021, doi: 10.1016/j.cose.2021.102324.

[10] M. L. Rodríguez-deArriba, A. L. Nocentini, E. Menesini, and V. Sánchez-Jiménez, "Dimensions and measures of cyber dating violence in adolescents: A systematic review," Aggress. Violent Behav., vol. 58, 2021, doi: 10.1016/j.avb.2021.101613.

[11] G. Y. Izadeen, "PRIVACY PRESERVATION SEMANTIC IN : WEB , INFORMATION , PARALLEL AND WEB COMPUTING," pp. 1523–1555, doi: 10.17605/OSF.IO/CBDRV.

[12] F. Khan, C. Ncube, L. K. Ramasamy, S. Kadry, and Y. Nam, "A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning,"

IEEE Access, vol. 8, pp. 119710–119719, 2020, doi: 10.1109/ACCESS.2020.3003785.

[13] J. Gu and S. Lu, “An effective intrusion detection approach using SVM with naïve Bayes feature embedding,” Comput. Secur., vol. 103, p. 102158, 2021, doi: 10.1016/j.cose.2020.102158.

[14] G. Liu, H. Zhao, F. Fan, G. Liu, Q. Xu, and S. Nazir, “An Enhanced Intrusion Detection Model Based on Improved kNN in WSNs,” Sensors, vol. 22, no. 4, pp. 1–18, 2022, doi: 10.3390/s22041407.

[15] and Q.-Y. Z. (Corresponding Wen-Tao Hao1, Ye Lu2, Rui-Hong Dong3, Yong-Li Shui3, “Adaptive Intrusion Detection Model Based on CNN and C5.0 Classifier,” vol. 24, no. 4, pp. 648–660, 2022, doi: 10.6633/IJNS.202207.

[16] M. S. Islam et al., “Using deep learning to detect digitally encoded DNA trigger for Trojan malware in Bio-Cyber attacks,” Sci. Rep., vol. 12, no. 1, pp. 1–13, 2022, doi: 10.1038/s41598-022-13700-5.

[17] M. D. Z. A. Farrukh Arslan, Aqib Javaid and A. Ebad-ur-Rehman, “Anomaly Detection in Time Series: Current Focus and Future Challenges,” Intech, vol. i, no. tourism, p. 13, 2023, doi: http://dx.doi.org/10.5772/57353.

[18] D. Farbiash and R. Puzis, “Cyberbiosecurity: DNA Injection Attack in Synthetic Biology,” 2020.

[19] F. Snigdha, S. M. M. Islam, O. Boric-Lubecke, and V. Lubecke, “Obstructive Sleep Apnea (OSA) Events Classification by Effective Radar Cross Section (ERCS) Method Using Microwave Doppler Radar and Machine Learning Classifier,” 2020 IEEE MTT-S Int. Microw. Biomed. Conf. IMBioC 2020, pp. 2020–2022, 2020, doi: 10.1109/IMBIoC47321.2020.9385028.

[20] M. Tawalbeh, M. Quwaider, and L. A. Tawalbeh, “IoT Cloud Enabeled Model for Safe and Smart Agriculture Environment,” 2021 12th Int. Conf. Inf. Commun. Syst. ICICS 2021, pp. 279–284, 2021, doi: 10.1109/ICICS52457.2021.9464567.

[21] R. Al Mogbil, M. Al Asqah, and S. El Khediri, “IoT: Security Challenges and Issues of Smart Homes/Cities,” 2020 Int. Conf. Comput. Inf. Technol. ICCIT 2020, pp. 258–263, 2020, doi: 10.1109/ICCIT-144147971.2020.9213827.

[22] A. M. Abdulazeez, M. A. Sulaiman, and D. Q. Zeebaree, “Evaluating Data Mining Classification Methods Performance in Internet of Things Applications,” J. Soft Comput. Data Min., vol. 1, no. 2, pp. 11–25, 2020, doi: 10.30880/jscdm.2020.01.02.002.

[23] A. Sabry Issa and A. Mohsin Abdulazeez Brifcani, “Intrusion Detection and Attack Classifier Based on Three Techniques: A Comparative Study,” Eng. Technol. J., vol. 29, no. 2, pp. 386–412, 2011, doi: 10.30684/etj.29.2.17.

[24] N. Sun et al., “Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives,” IEEE Commun. Surv. Tutorials, vol. 25, no. 3, pp. 1748–1774, 2023, doi: 10.1109/COMST.2023.3273282.

[25] S. Silvestri, S. Islam, S. Papastergiou, C. Tzagkarakis, and M. Ciampi, “A Machine Learning Approach for the NLP-Based Analysis of Cyber Threats and Vulnerabilities of the Healthcare Ecosystem †,” Sensors, vol. 23, no. 2, pp. 1–26, 2023, doi: 10.3390/s23020651.

[26] A. Mozo, A. Karamchandani, L. de la Cal, S. Gómez-Canaval, A. Pastor, and L. Gifre, “A Machine-Learning-Based Cyberattack Detector for a Cloud-Based SDN Controller,” Appl. Sci., vol. 13, no. 8, 2023, doi: 10.3390/app13084914.

[27] M. Abushwereb, M. Mustafa, M. Al-kasassbeh, and M. Qasaimeh, “Attack based DoS attack detection using multiple classifier,” no. Mid, 2020.

[28] B. H. Krishna and D. K. Sattivada, “Big Data Cyber Security Using Machine Learning,” Europeanchemicalbulletin, vol. 12, no. s issue4, pp. 11074–11082, 2023, doi: 10.48047/ecb/2023.12.si4.1001.

[29] S. Toliupa, S. Buchyk, V. Nakonechnyi, V. Saiko, I. Parkhomenko, and N. Lukova-Chuiko, “Building an Intrusion Detection System in Critically Important Information Networks with Application of Data Mining Methods,” Proc. - 16th Int. Conf. Adv. Trends Radioelectron. Telecommun. Comput. Eng. TCSET 2022, pp. 128–133, 2022, doi: 10.1109/TCSET55632.2022.9767029.

[30] R. K. Dubey, N. Dandotiya, A. Sharma, S. Mishra, and S. K. Gupta, “Cyber attack Detection Using Machine Learning Techniques,” 3rd IEEE Int. Conf. ICT Bus. Ind. Gov. ICTBIG 2023, pp. 1–6, 2023, doi: 10.1109/ICTBIG59752.2023.10456080.

[31] A. Cuzzocrea, E. Fadda, and E. Mumolo, “Cyber-attack detection via non-linear prediction of IP addresses: an innovative big data analytics approach,” Multimed. Tools Appl., vol. 81, no. 1, pp. 171–189, 2022, doi: 10.1007/s11042-021-11390-1.

[32] B. Murugeshwari, D. Selvaraj, K. Sudharson, and S. Radhika, “Data Mining with Privacy Protection Using Precise Elliptical Curve Cryptography,” Intell. Autom. Soft Comput., vol. 35, no. 1, pp. 839–851, 2023, doi: 10.32604/iasc.2023.028548.

[33] N. Mohamed and M. M. A. Salama, “Data Mining-Based Cyber-Physical Attack Detection Tool for Attack-Resilient Adaptive Protective Relays,” Energies, vol. 15, no. 12, 2022, doi: 10.3390/en15124328.

[34] M. Ahsan, R. Gomes, M. M. Chowdhury, and K. E. Nygard, “Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector,” J. Cybersecurity Priv., vol. 1, no. 1, pp. 199–218, 2021, doi: 10.3390/jcp1010011.

[35] J. A. Mata-Torres, E. Tello-Leal, J. D. Hernandez-Resendiz, and U. M. Ramirez-Alcocer, “Evaluation of Machine Learning Techniques for Malware Detection,” Intell. Syst. Ref. Libr., vol. 226, pp. 121–140, 2023, doi: 10.1007/978-3-031-08246-7_6.

[36] S. Jorgensen et al., “Extensible Machine Learning for Encrypted Network Traffic Application Labeling via Uncertainty Quantification,” IEEE Trans. Artif. Intell., vol. 5, no. 1, pp. 420–433, 2024, doi: 10.1109/TAI.2023.3244168.

[37] N. A. Majedkan, B. A. Idrees, O. M. Ahmed, L. M. Haji, and H. I. Dino, “Queuing Theory Model of Expected Waiting Time for Fast Diagnosis nCovid-19: A Case Study,” 3rd Int. Conf. Adv. Sci. Eng. ICOASE 2020, no. June 2021, pp. 127–132, 2020, doi: 10.1109/ICOASE51841.2020.9436601.

[38] H. T. Sadeeq and A. M. Abdulazeez, “Metaheuristics: A Review of Algorithms,” Int. J. online Biomed. Eng., vol. 19, no. 9, pp. 142–164, 2023, doi: 10.3991/ijoe.v19i09.39683.