Predictive AI for Identifying Undiscovered Cyber Threats: A Proactive Security Model for Big Data

Asmaa Ali Jasim[1], Mohammed Hasan Hadi [2]

[1] *Open Educational College, Ministry of Education , IRAQ*

[2] *Open Educational College, Ministry of Education , IRAQ*

[*]*Corresponding Author:* **mohammed.almaawi.iq@gmail.com**

Received 11 Dec. 2024, Accepted 7 May. 2025, Published 30 June. 2025.

DOI: 10.52113/2/12.01.2025/156-175

Abstract: As technology advances, cybercriminals adopt more sophisticated strategies to attack weaknesses in individual computers, organisational networks, and nation-states. Organisations systematically gather substantial quantities of security-relevant data, including log events from individuals, networks, and software applications, for further forensic analysis. Conventional security analysis methods are inadequate for handling huge data volumes and may generate excessive false alarms, particularly when organisations transition to cloud architectures and accumulate more data. Furthermore, the identification of current and more complex assaults, such as persistent and advanced threats (APTs), requires ongoing monitoring and analysis of extensive security-related data, with precision and speed. Big Data analytics is actively used in several domains, including financial transactions, healthcare, and industrial applications, among others. It has recently garnered the interest of the information security community because to its purported capability to correlate security-related data and derive insights effectively at an unprecedented scale. In this study, we examine the limitations of conventional technology/systems and SIEM tools in handling massive amounts of data and complex, advanced threats. We further examine the prerequisites for the effective use of Big Data analytics in the domains of cyber threat intelligence and cybersecurity to address extensive data volumes and complex threats. Ultimately, we emphasise the issues arising from this adoption and provide solutions to address these challenges in future study.

Keywords: Cyber Threats, A Proactive Security Model, Big Data Analytics, Big Data.

1.Introduction

Merriam-Webster defines cyber-security as measures used to protect a computer or computer system, particularly on the Internet, against unauthorised access or attacks. The United States National Initiative for Cybersecurity Careers and Studies (NICCS) offered a comprehensive definition of cybersecurity. Cybersecurity is described as the strategy, policy, and standards pertaining to the security of and activities in cyberspace, including the whole spectrum of threat reduction, vulnerability mitigation, deterrence, international collaboration, incident response, and resilience. [1].

It also encompasses recovery policies and practices, including computer network operations, information assurance, law enforcement, diplomacy, military [2]. There were 1,966,324 reported instances of malware infections in 2015 that attempted to hack into online bank accounts in order to steal money, according to the most recent numbers provided by Kaspersky in their final statistic report [2]. Just looking at just one economic sector reveals the enormous threat that

malware infections pose to the global economy annually [3].

Given the exponential growth of cyberattacks, it is evident that current measures to protect IT infrastructure, company networks, and online applications may be insufficient. Next, we need to ask: how can we better defend ourselves and identify the exponential rise of these cyberattacks? The US federal government, energy, and financial services were the subjects of an interview by International Data Corporation (IDC) in 2015 with security experts, executives, and specialists from each area. Gaining insight into how cyber dangers have changed over time was the primary goal of the interview [4].

The interviews led to the conclusion that cyber-security dangers are on the rise and that businesses should stop responding to security incidents after the fact and start proactively assessing risks before they can do harm [3]. The volume of daily events generated by major organisations may reach 100 billion, according to [5]. Any growth in data sources, personnel, equipment, or software applications will inevitably lead to a rise in the total number of events. Cybersecurity tools including intrusion detection systems, log events analysis, and others rely on antiquated methods that fail to scale, generate excessive false alarms, and are otherwise unsuitable for the job [4].

As more and more data is collected by the organisation via cloud infrastructures, the situation becomes worse [6]. Thus, cyber threat intelligence is necessary to provide the continuous real-time gathering and monitoring of data streams, which in turn allows for the activation of an appropriate risk mitigation procedure prior to assaults causing significant damages [7].

The term "security information and event management" (SIEM) refers to a system that helps with security operations and analysis by collecting and analysing data about network flows, security incidents, and logs [5]. Cyber threat intelligence incorporates features that have been used in security information and event management systems, such as the ability to manage logs, correlate security events, and monitor network activities [6].

Most organisations use SIEM to monitor risks instead of doing conventional security analysis and investigations, according to [5]. Having said that, SIEM isn't robust enough to handle the massive evolution of threats in the modern day. These deficiencies, as stated in [5], include the following: One issue is that new assaults utilise multidimensional strategies and vary from system to system. SIEM's event correlation depends on data that is normalised in respect to predetermined schemas, making it difficult to adjust [7].

Additionally, SIEM-based solutions have a hard time keeping up with the ever-increasing volume of events since they use fixed storage (Schema). Additionally, Third, SIEMs are built on top of predetermined context, making them situationally particular and requiring the time-consuming process of re-definition before they can be applied to various scenarios; Fourth, adopting new rules necessitated reconstructing the whole method because to SIEMs' inflexibility. Enterprises need a fresh strategy for cyber-security that addresses these issues, according to the authors in [5]. Information security team expertise and Big Data-based security analytics-technologies establish an end-to-end interaction, according to the new method.

2.Literature Review

Big Data Analytics, a comprehensive information processing and analysis methodology, has been actively used across several domains. It also captivates the attention of the information security community due to its promising capability in efficiently analysing and correlating security-related data [4]. Authors in [8] indicated that novel and previously unobserved cyberattacks are on the rise owing to deficiencies in current

security mechanisms. Threats range from the dissemination of personal information to service disruption and assaults against extensive systems, including vital infrastructure [7].

It is found that the majority of unidentified cyber threats are overlooked because standard security technologies rely on basic pattern matching and must be addressed within the framework of Big Data analytics. Limited Big Data analytics-based security solutions were identified in the literature, including references [8-12]. Authors in [9] examined the deployment of cyber-security insurance (CI) in cloud-based services and developed a safe architecture for cyber event analytics using Big Data.

The authors said that their methodology was developed for aligning various cyber risk scenarios, using repository data. The simulation results have shown the theoretical validity of the framework's adoptability and practicality, as asserted by the authors. In [10], the writers were inspired by the smart grid, a promising system capable of meeting renewable energy needs via the integration of modern information and communications technology (ICT). Consequently, the authors asserted that the widespread implementation of advanced ICT would provide substantial energy data characterised by volume, velocity, and diversity, particularly via the use of smart metering [11].

The development of Big Data offers significant advantages for enhanced energy conservation, planning, and efficient energy generation and distribution; yet, it also introduces new security concerns related to user privacy and the safe functioning of key infrastructure. The produced Big Data may provide significant advantages for improved energy planning, energy-efficient production, and delivery. Nonetheless, these privacy and security concerns must be addressed. Consequently, the authors examined smart technologies for renewable energy sources and associated Big Data security concerns [12].

In [10], the authors provided a comprehensive study of the current advancements in security analytics, including its definition, technology, patterns, and tools. Ultimately, the writers hoped their work would enlighten readers about analytics' potential future use as an unrivalled cyber-security solution. If you want to safeguard your system against APTs, read the authors' proposal in [11]. The proposed architecture integrates several methodologies grounded in Big Data analytics including security intelligence to assist human analysts in prioritising hosts with the highest likelihood of penetration.

In [12], the authors introduced a Cyber Security analytics framework designed for thorough cyber security monitoring by integrating cyber security-related events with feature selection to predict user behaviour based on diverse sensors. The suggested framework for Cyber Security Analytics (CSA) utilises Big Data analytics and is predicated on Network Log (NetL) and in-memory Process Log (PrcL) to detect an abnormality vector via extensive system observations [13].

The transition from traditional security solutions to Big Data-based security systems for long-term, large-scale analytics occurred for three primary reasons, as stated in [14]: Initially, managing substantial data volumes was not economically viable under conventional security systems and SIEM; hence, some event recordings were often purged after a certain period to free up storage for new occurrences. Second: the challenge of executing data analytics and intricate queries on extensive and diverse datasets characterised by noisy and missing attributes. Third: managing huge data warehouses is costly, since their implementation need robust business justifications [15].

3. Cybersecurity and Cyber Threat Intelligence

Cybersecurity is described as a collection of countermeasures, methods, and standards used to prevent, detect, and protect against vulnerabilities inside systems, organisational networks, or the Internet in cyberspace. Attack: "An endeavour to unlawfully access system resources, services, or information, or to undermine system integrity." 1. Advanced Persistent Threat (APT): “An opponent endowed with advanced skill and substantial resources, enabling it to use several attack vectors (e.g., cyber, tangible, and deception) to fulfil its objectives.” [8].

According to the Gartner dictionary, threat intelligence is defined as "evidence-based knowledge that encompasses context, techniques, indicators, implications, and actionable recommendations concerning an existing or emerging threat to assets, which can inform decisions related to the subject's response to that threat." Security Information and Event Management (SIEM) is characterised by the Enterprise Strategy Group (ESG) as “a platform designed to aggregate and correlate security events, logs, and network flow data for the purposes of security analysis and operations.” [11].

3.1 Conventional Cybersecurity Analysis and Management Strategies

Different security analysis and management techniques and processes have been utilised to defend and safeguard IT systems ever since the idea of cyber-security was first introduced. It is used to safeguard organisational networks and the Internet from many forms of cybersecurity attacks. IDC categorises these cyber-security risks into ten overarching categories, as seen in Figure (1). All risks encountered by systems and networks today fall into these categories or their derivatives. Distributed denial of service attacks (DDoS), advanced persistent threats (APTs), and zero-day assaults are the most complex and protracted threats that need timely and precise detection.

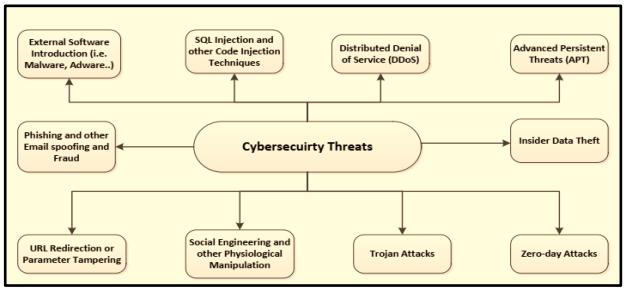


Fig. (1): Illustrate the Cybersecurity Threats

Throughout the history of information systems and network security, several techniques and approaches have been developed and advanced to protect against and alleviate the impacts of cybersecurity threats. Figure (3, 2) illustrates the conventional methodologies and procedures for cybersecurity management and analysis often used in enterprises or individual IT systems. A concise overview of each method is provided in the following paragraphs.

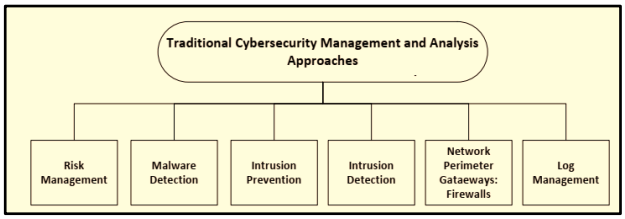


Fig. (2): Conventional Cybersecurity Management and Analytical Methods.

3.2 Risk Management

In the context of business, a risk was defined as the potential for an event to diminish the value of the business. The incident is sometimes referred to as a "adverse event." The authors in [13] contended that information security involves information risk management as well. Furthermore, they said that to assess the dangers and efficacy of security risk mitigation strategies in the realm of information security, some critical data must be gathered [12].

This information encompasses potential vulnerabilities in the information security system, data pertaining to global business security incidents, the direct and indirect losses incurred from each incident, and the

countermeasures employed to mitigate such incidents arising from these vulnerabilities. Several risk management frameworks and methodologies have been developed in the literature on information security [14].

A recent technique for information security risk analysis was described in [14], using fuzzy decision theory and event tree analysis. The model recognises and assesses the sequence of events in an incident scenario subsequent to the possible misuse of information technology systems. The research in [15] introduced a hybrid technique for information security risk assessment that integrates both quantitative and qualitative risk management methodologies.

The authors evaluated the merits and drawbacks of both analytical methods and found that no method can attain optimal performance independently. Consequently, a hybrid methodology that employs precise decisions from the quantitative approach may be integrated with the qualitative approach grounded on judgements and intuitions [16].

4.Model for Big Data

Big Data has emerged as a critical area of contemporary and future study. Gartner [47] identified the transformation of Big Data into ubiquitous intelligence as one of the three components of the top 10 key technology trends during 2015. Gartner defines Big Data as “high-volume, high-velocity, and/or high-variety data assets that need cost-effective, creative methods of information processing to provide improved insight, decision-making, and process automation.” A different description was provided in [17] as an extensive and varied collection of data sets that is challenging to analyse using both conventional and modern data processing tools. Figure 3 illustrates the three aspects of Big Data and their respective views. Additional studies, such as those in [18], have included a fourth "V" into the concept, denoting the value of data.

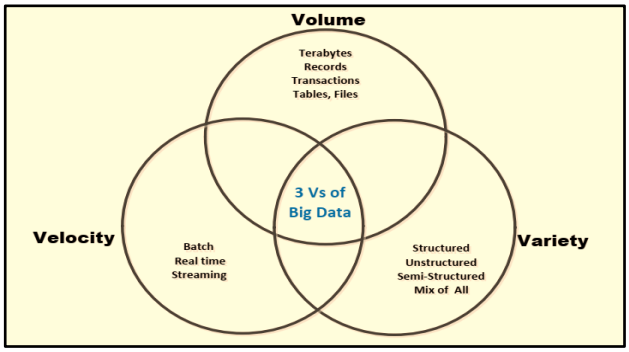


Fig. (3): The three "Vs" of big data are volume, velocity, and variety.

The initial "V" denotes volume, signifying the generation of terabytes of data records from transactions, tables, and files. Massachusetts reports that Facebook produces over 500 terabytes of data each day. In 2014, it was stated that Facebook's data warehouse had a capacity of 300 PB, with an incoming daily data flow of about 600 TB. The second "V" denotes velocity, indicating that data is characterised by real-time streaming created at an exceedingly rapid pace. Massachusetts addresses Big Data issues by analysing 2 million records every day to determine the causes of certain data losses. The third "V," denoting variety, illustrates the incorporation of diverse data types from multiple sources in Big Data, including structured data (e.g., employee records in an organisation), unstructured data (e.g., images, audio, video, sensor data), semi-structured data, or a combination of these types concurrently [16].

4.1 Applications of Big Data

Diverse application domains provide as sources of data Big Data analytics. Applications include, but are not limited to, education, scientific fields, retail, history, cultural activities, government, healthcare, social networking, finance, and transportation. Figure (4) illustrates the many application domains. A concise overview of the implementation of Big Data analytics in certain important application domains is provided in the following paragraphs. This

evaluation does not provide a comprehensive explanation of all application areas [18].

Big Data analytics, as defined in is the application of sophisticated analytical methods to large data sets. Consequently, Big Data analytics integrates two concepts: Big Data and analytics. It further integrates the manner in which the two concepts might be combined to create one of the most prominent paradigms in business intelligence (BI) today. The following paragraphs succinctly examine several application domains of Big Data analytics across many facets of human existence. For every app, we demonstrate the need of embracing the Big Data analytics approach [19].

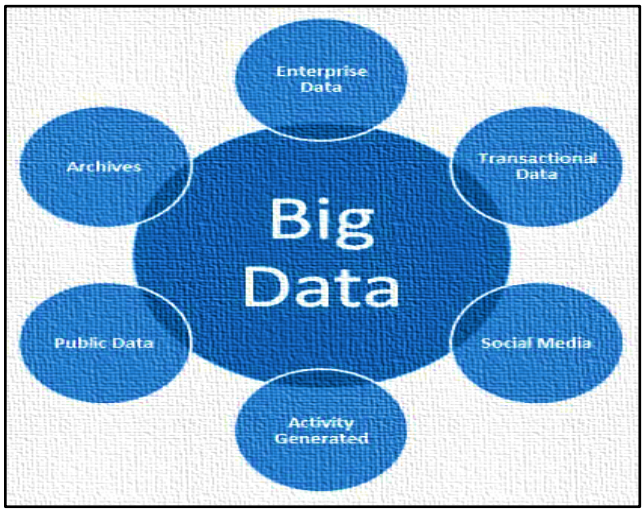


Fig. (4): The primary sources of big data are social, machine, and transactional [13].

4.2 Big Data Analytics Tools and Techniques

A specialised processing framework is required to handle Big Data. In 2004, Google introduced MapReduce as a programming methodology designed for the generation and processing of massive data sets using cluster-based parallel and distributed algorithms. MapReduce has two primary processes: the map process and the reduce process. During the mapping process, a collection of intermediate key/value pairs is produced using a user-specified map function [17].

During the reduction phase, all intermediate value pairs linked to the same intermediate key are consolidated using a reduction function. Open-source software for dependable, scalable, and distributed computing, Hadoop revolves on MapReduce. The Apache Hadoop software library is a platform that facilitates the distributed processing of large data sets on computer clusters using programming paradigms like MapReduce. This program is designed for scalability, accommodating configurations from a single computer to thousands, with each machine possessing its own local resources [15].

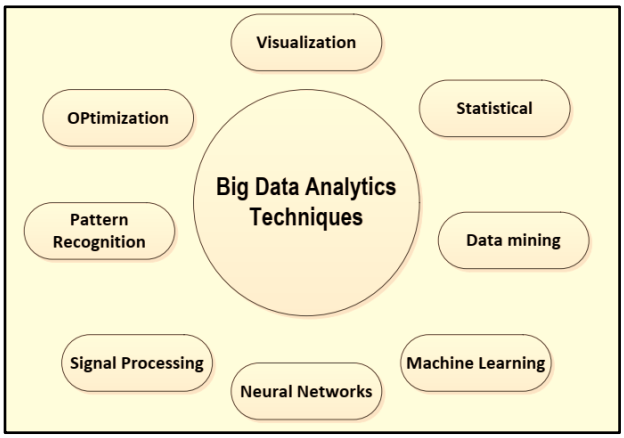
The service availability of Hadoop is predicated on its architecture to identify potential faults at the application layer rather than depending on the hardware in the underlying layers. Consequently, the service can be consistently provided atop a cluster of computers. The Hadoop project has four primary modules: Hadoop Common, Hadoop Distributed File System, Hadoop YARN, and Hadoop MapReduce. Numerous Hadoop-related projects provide frameworks for distributed and parallel computing, as documented in the literature, including Ambari, Avro, Cassandra, Chukwa, HBase, Hive, Mahout, Pig, Spark, Tez, and Zookeeper. For more information on these systems, the reader may see [18].

As stated in [49], current Big Data solutions concentrate on three processing categories: batch processing, stream processing, and interactive analysis. Tools like Mahout and Dryad, which rely on batch processing, often make use of Apache Hadoop infrastructure. Applications that work with stream data in real time need the stream processing tools. Storm and S4 exemplify distributed and parallel streaming data analytics solutions. The interactive analysis allows users to analyse their data in real-time [19].

Each user connects to the computer and may engage interactively in an online format. The user may examine, contrast, and evaluate the data in tabular or graphical formats, or both simultaneously. Exceptional processing and

analytical methodologies are required for Big Data to evaluate vast data volumes within limited timeframes. Every Big Data application necessitates an appropriate Big Data approach, as stated in [17].

For example, online retail companies like Wal-Mart use machine learning and statistical methodologies to identify trends in their transactional data, so enhancing their



competitive edge in pricing tactics and advertising. Multiple disciplines are engaged in Big Data processing and analytical methodologies owing to the concept's application across several sectors. Figure (5) illustrates the main and prevalent disciplines for Big Data processing. These fields include statistics, data mining, machine learning, neurocomputing, signal processing, pattern recognition, optimisation, and visualization [20].

Figure (5): Techniques and Disciplines of Big Data [21].

A comprehensive examination of Big Data methodologies is available in. Numerous statistical methodologies, including cluster analysis, factor analysis, correlation analysis, and regression analysis, have been used in conventional processing systems. They may also be adapted for Big Data processing by optimising them for high-performance computing [19]. Likewise, data mining methodologies such C4.5, k-means, SVM, Naïve Bayesian, and Belief Bayesian, among others, have been used for data processing in conventional non-Big Data systems. The primary consideration in Big Data processing is the extraction of significant information and insights from extensive datasets to derive conclusions beneficial for organisations and individuals.

4.2.1 Machine Learning Techniques

Machine learning (ML) is a subset of artificial intelligence (AI) that focuses on building systems that can learn from data and improve their performance over time without being explicitly programmed. There are various machine learning techniques, which can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning. Here's an overview of these techniques and some common algorithms within each category:

1. Supervised Learning

Supervised learning involves training a model on labeled data, where the input data (features) and the corresponding output (target) are provided. The goal is to learn a mapping from inputs to outputs. It consist of the following :

Common Techniques: Regression: Predicts continuous values. Linear Regression. Polynomial Regression. Ridge Regression. Lasso Regression

Classification: Predicts discrete labels. Logistic Regression. Decision Trees. Random Forests. Support Vector Machines (SVM). k-Nearest Neighbors (k-NN). Naive Bayes. Neural Networks

Applications: Predicting house prices (regression). Spam detection (classification) . Image classification

2. Unsupervised Learning

Unsupervised learning involves training a model on unlabeled data, where the goal is to find hidden patterns or structures in the data.

Common Techniques: Clustering: Groups similar data points together. k-Means Clustering**.** Hierarchical Clustering**.** DBSCAN**.** Gaussian Mixture Models (GMM)**.** Dimensionality Reduction: Reduces the number of features while preserving important information. Principal Component Analysis (PCA)**.** t-Distributed Stochastic Neighbor Embedding (t-SNE)**.** Uniform Manifold Approximation and Projection (UMAP)**.** Anomaly Detection: Identifies unusual data points. Isolation Forest**.** One-Class SVM [20].

Applications: Customer segmentation (clustering)**.** Feature extraction for visualization (dimensionality reduction)**.** Fraud detection (anomaly detection)

3. Reinforcement Learning

Reinforcement learning involves training an agent to make decisions by rewarding desired behaviors and punishing undesired ones. The agent learns by interacting with an environment.

Common Techniques: Model-Based RL: Uses a model of the environment to plan actions. Dynamic Programming**.** Model-Free RL: Learns directly from interactions without a model. **.**Q-Learning**.** Deep Q-Networks (DQN)**.** Policy Gradient Methods**.** Actor-Critic Methods

Applications: Game playing (e.g., AlphaGo). Robotics. Autonomous vehicles

4. Semi-Supervised Learning: Semi-supervised learning combines labeled and unlabeled data to improve learning accuracy. It is useful when labeled data is scarce.

5. Deep Learning

Deep learning is a subset of machine learning that uses neural networks with multiple layers to model complex patterns in data.

6. Ensemble Learning

Ensemble learning combines multiple models to improve performance and robustness.

7. Transfer Learning

Transfer learning involves using a pre-trained model on a new, related problem. It is especially useful when data is limited.

8. Natural Language Processing (NLP)

NLP focuses on enabling machines to understand and process human language.

9. Time Series Analysis

Time series analysis involves analyzing sequential data points collected over time.

10. Anomaly Detection

Anomaly detection identifies rare or unusual data points that deviate significantly from the norm [19].

4.3 Fundamental Challenges of Big Data

Despite the allure and potential of Big Data analytics across several domains, several problems impede its rapid proliferation. Several of these problems were examined in [66] and may be succinctly summarised in the following subsections. We classified these difficulties into three primary categories [28]: Big Data administration, visualisations, and security and privacy.

1.Challenges Associated with Data Management: Four primary issues distinguish Big Data management from conventional data management. The problems are to data warehousing, data variation, integration of data, and data processing as well as resource management. In the context of Big Data warehousing, Big Data is often stored as vast quantities of unstructured data aggregated from many sources. The problem lies in efficiently storing and extracting significant information from the enormous volume of unstructured data [32].

What is the optimal method for storing such data to ensure timely retrieval? Is the existing file system technology enough for storing Big Data, and what enhancements are necessary to optimise its suitability? What techniques should be used for migrating Big Data across data centres or cloud providers? What is the level of transparency of these tactics from the user of Big Data? How can one manage

extensive unstructured data from diverse sources in the context of Big Data diversity? How can one efficiently extract pertinent extracts from extensive data? What is the optimal method for aggregating and correlating the retrieved data to get useful insights and conclusions? Regarding Data Integration: does Big Data necessitate the development of new protocols and interfaces for handling diverse data kinds from many sources?

Ultimately, regarding data processing or resource management, is there a need to develop new programming models to address streaming and multidimensional data? How may one enhance resource utilisation, particularly energy usage, in streaming data system applications like wireless sensor networks? All these problems must be addressed during the planning phase of Big Data analytics [19]. All application domains have similar issues, but with varied degrees of difficulty from one application to another.

2.Challenges Associated with Big Data Visualisation: Efficient Big Data processing algorithms are essential for real-time visualisation of Big Data. Authors observed that several computational techniques used for Big Data analytics are intricate and need meticulous parameter modification to adapt to specific real-time scenarios. However, doing such actions may be essential and time-intensive. The authors determined that certain strategies must be used in the realm of human-machine interactions to provide efficient and timely data visualisations [18]. These strategies encompass: 1- reducing the accuracy of outcomes. 2- diminishing the convergence within the computational model. 3- limiting data scale. Four data points undergoing coarse processing. 5- according to the resolution limitations of the visualisation equipment. Authors emphasised the significance of visualisation in the management of computer networks or software analytics, particularly in relation to large-scale infrastructure data analytics [34].

3.Challenges Concerning Big Data Privacy and Security: Separating Big Data security from security that makes use of Big Data ideas is a good idea. The first one is about making sure Big Data is safe from security breaches of any type, and the second one is about using Big Data ideas to make cyber defences stronger. This part provides a high-level overview of Big Data privacy and security concerns, which are among the most pressing problems with Big Data; part 4 delves more into the topic of using Big Data analytics to strengthen cyber defences [29].

There is a delicate balance between lawful data usage and consumer privacy, according to the academics that examined Big Data's security and privacy concerns. So, from a security perspective, the most pressing issue with Big Data is how to ensure the privacy of its users. This is due to the fact, as the author points out, that security breaches using Big Data may have far-reaching and multi-faceted effects, making them much more disastrous than those involving regular data. In order to stay in compliance, the author recommended that businesses identify the most sensitive parts of Big Data, including customer IDs, and securely remove them. The old method of organising massive datasets made it easy to obtain and query any sensitive information [8]. Nevertheless, sensitive data queries become more complicated and time-consuming when dealing with Big Data, which includes a mix of data that is structured, unstructured, and semi-structured. From a security perspective, it is possible that different users may need access to different subsets of data. In order to accommodate such requests, it will be necessary to adapt existing encryption and access control technologies to meet the needs of the new data format. For instance, it is important to ensure that only authorised users

may access data by designing the access control policy accordingly [11].

As mentioned in, there are five main security-related factors to consider when working with Big Data: data encryption, data anonymisation, access control, security policy, monitoring, & governance structures. To solve privacy issues, anonymisation removes sensitive information from Big Data. while it comes to encryption, the idea is to use encrypted data rather than plaintext data while doing operations. To further guarantee security and privacy preservation, real-time threat intelligence and Big Data monitoring are also essential [23].

Finding sensitive information in unstructured data is a policy concern. As an added precaution, it deletes the data from storage the moment it is no longer required. The novelty of the Big Data idea has delayed the development of relevant rules and procedures in the realm of governance systems, but new approaches are beginning to emerge to address this challenge [18]. There is a serious, maybe fatal, problem with healthcare data privacy and security. That is because protecting patients' privacy is just as important as following all applicable regulations and protocols when dealing with their medical records. Mishandling certain patients' data might have fatal effects. A number of pharmacological datasets are also significant intellectual property that need safeguards even in controlled settings [24].

4.4 Adoption of Big Data Analytics for Cyber-security

Data-centric information security has been used for decades in the realms of bank fraud detection and anomaly-based intrusion detection systems. Due to the massive amounts of data produced by fraud detection systems—millions of data instances and events each day—for medium and large businesses, these systems qualify as Big Data analytics solutions. Fraud detection is used in several sectors, including credit card firms, healthcare, insurance, and telecommunications, among others [23].

Intrusion detection within the realm of data analytics has progressed through three generations, as seen in Figure (6).

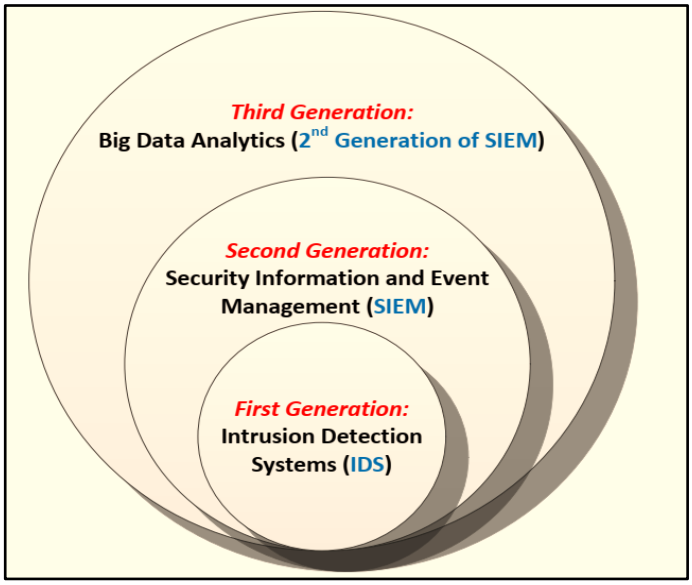


Fig. (6): Advancement of Intrusion Detection within the Framework of Data Analytics [31].

As discussed in Section 2, the first version was developed to detect security breaches that were overlooked by other tools for protecting the perimeter of a network, such as risk assessments and malware detectors. Generation 2: SIEM, or security information and event management, adds an extra rule for collecting and sorting alerts from various first-generation sources (i.e., intrusion detection systems) [24]. Third generation: Big Data security analytics, often regarded as second generation SIEM. It enhances the initiatives established in the second generation by minimising the time required for correlating and aggregating security incident data. Furthermore, it utilises contextual and long-term historical data for the forensic study of cyber threats [19].

4.5 Requirements of Big Data Analytics for Cyber-security

Any Big Data analytics-based cybersecurity solution must meet certain criteria to address the ongoing and fast escalation of complex cyber threats [10]. These criteria must account

for the inherent properties of Big Data with the security considerations. Our analysis indicates that the following set of conditions must be taken into account [32].

1. Management of Data at a Large Scale:

If cyber-security solutions based on Big Data analytics are to keep up with the ever-increasing sophistication of cyber-attacks, they must meet certain criteria. These standards must account for the attributes of Big Data with the security necessities [22]. Our analysis indicates that the following set of conditions must be taken into account.

A. Managing Multi-Sourced Data: There is significant expansion in data sources available for cyber-security systems within the framework of Big Data analytics. Sources include, but are not limited to, firewall logs, directory access files, operating system event logs, SIEM data, Dpi data, SQL server logs, NetFlow data, threat analysis data, and more resources. Such data sources have existed for a considerable duration but were not collectively used in the framework of Big Data analytics [24]. Integrating data from many sources is essential for obtaining valuable insights to more effectively identify and prevent cyber-attacks [33].

B. Manage data on a grand scale: The design of cyber-security systems based on Big Data analytics is becoming more challenging due to the growing data quantities caused by the proliferation of data sources. So, to efficiently gather, analyse, and retrieve relevant data in a timely way, this need should be considered throughout the design of such systems. If you are building a cyber-security system around big data, you should make good use of cloud computing, clustering, and grid computing as platforms to store, analyse, retrieve, and derive insights from data when you need them [27].

C.

D. Working with Different Kinds of Data: With more and more data sources popping up, it is becoming more common to meet a wide range of data kinds, from very organised to extremely unstructured information. Most of the data used in the pats was numerical and came from a single data type, which was helpful for security analytics. These days, nevertheless, it is possible to collect very unstructured data from a wide variety of sources, including, but not limited to, e-mails, blog posts, social network activity, threat feeds, and more [28].

2. The visualisation of data: Visualisation It is a crucial element in cyber threat intelligence which provides a graphical descriptive evaluation of security-related data. The visualised relationships among devices, events, places, signatures, and IP addresses facilitate the identification of data abnormalities and intrusions. Consequently, visualisation is essential for comprehending these relationships and deriving insights about the behaviour of network systems [29].

A. Keylines, a visualisation dashboard created by Cambridge Intelligence Corporation, is a network visualisation tool intended to depict cyber hazards, enabling users to conduct more efficient and effective data analysis. It offers a method to get significant insights from intricate interconnected cyber data. It has four primary capabilities: (i) analysing software risks and weaknesses; (ii) detecting unusual logins; (iii) identifying trends in data breaches; plus (iv) monitoring malware dissemination patterns over time. The significance of these visualisation apps rests in their ability to engage users in identifying patterns and anomalies by transforming raw linked data into dynamic interactive charts [30].

B. The need for visual analytics was highlighted as a tool that assists security teams in comprehending relationships and monitoring historical trends among security data pieces. The Visual Analytics Suites for Cyber Security (VACS) is a visual analytics system that integrates multi-criteria clustering approaches and employs three forms of interactive visualisations: treemaps, node-link diagrams, and chord diagrams. The VACS, like Keylines, sought to get insights from diverse threat environments. VACS was primarily developed as a dashboard interface that offers an overview of host-based thumbnails and facilitates querying and retrieving information to analyse questionable hosts [31].

3- Infrastructure Technology for High Performance: It is essential to conduct a comprehensive study of the fundamental infrastructure technologies that underpin Big Data analytics, including online computing, distributed computing, computational grids, stream processing, Big Data modelling, Big Data architecture, and software systems. In cybersecurity, like in other Big Data applications, it is evident that the data used as proof of assaults and security breaches is expanding across the three dimensions of Big Data: volume, velocity, and diversity. This development complicates the detection of such assaults using conventional techniques [34]. Detecting the most sophisticated APT assaults just via conventional information retrieval systems used in standard IDS architecture is challenging. Instead, modern technologies like as the MapReduce framework should be used. Utilising a MapReduce implementation enhances APT detection systems' capacity to effectively manage complex unstructured data, which varies in format and is sourced from diverse origins such as system logs, IDS, NetFlow, as well firewalls, and DNS systems over extended durations [35].

Furthermore, MapReduce's capacity for extensive parallel processing enables the implementation of very advanced detection algorithms that conventional SQL-based data systems cannot accommodate. The MapReduce architecture, with map and reduce functions, facilitates user flexibility in integrating additional detection techniques. This approach makes the distributions apparent to users engaging directly with individual data. It may be inferred that the deployment of large-scale distributed computing systems will facilitate the concurrent analysis of vast data volumes, hence offering a means to identify additional attack vectors and targets for the detection of unknown and complex threats such as APTs [36].

4.6 Challenges of Big Data Analytics adoption in Cybersecurity

Variety is one of the aspects that characterises Big Data. Big Data analytics systems are characterised by several forms of data, including structured, semi-structured, and unstructured data. Conventional security analysis solutions, including log mining, intrusion detection, and SIEM systems, manage well-structured data derived from a singular data source, such as system log files, database logs, or Netflow records. Nonetheless, the integration of unorganised or semi-structured data from diverse information sources, including emails, social media, threat feeds, and other security-related materials, alongside established structured security data, remains a difficulty [37].

Two primary variables influence the management of unstructured data: the velocity of data development and the proliferation of data sources. Consequently, addressing the challenge of unstructured data necessitates the simultaneous consideration of both elements. Big Data analytics for cybersecurity systems must be engineered to accommodate the rapid

expansion of data sources and the substantial volume of data collected over time [38].

4.6.1 Real Time Analysis

Streaming data analysis entails the rapid processing of substantial volumes of data in near real-time. It is sometimes referred to as processing data in motion, which is directly related to real-time analytics mentioned above. Conventional security analysis systems analysed data streams in a batch fashion, wherein historical data was analysed after a designated interval. An illustration of this is the procedure for identifying fraud in financial transactions. The data for hours or even a day is gathered and analysed at the conclusion of that time frame [32].

This analysis must be rapid and conducted in real time to be effective and to provide timely, proactive responses before more harm occurs due to malicious acts. Security-related data has a streaming characteristic owing to the dynamic nature of network architecture. Moreover, the progress of infrastructure is rapid, particularly with Internet of Things (IoT) applications that rely on sensor technology. Consequently, this streaming analysis capability poses a barrier for Big Data analytics when the application necessitates real-time analysis and reaction [39].

4.6.2 Visual analytics

It is the process of using interactive visual interfaces to analyse information for people in order to get further knowledge and insights from the presented topic. This research previously examines the significance of visual analytics in cyber threat intelligence. Keylines was presented as an example of cyber threat intelligence visualisation tools that may be used with cybersecurity dashboards to display network connection and generate reports on the network's status at a specific moment in time. Visual analytics poses a problem for Big Data analytics-based security solutions, particularly for the real-time analysis of streaming data [40].

The visualisation dashboard must be constructed to monitor the variability of events produced by motion while offering security analysts with previously inconspicuous insights, such as atypical incoming or outgoing traffic of a host or group of hosts inside a network segment. The complexity of modelling network connections, particularly in large-scale companies, exacerbates the challenges faced by visual analytics in Big Data cybersecurity systems [38].

4.6.3 Data Confidentiality

A significant impediment to the effective use of Big Data analytics in cybersecurity is data privacy. Privacy concerns contravene the concept of reuse, which stipulates that shared data should be used only for its intended objectives. In the realm of conventional data utilisation, privacy pertains to the implications arising from the management of sensitive datasets. Nonetheless, Big Data analytics complicates privacy violations, since it enables the extraction of insights and conclusions on people or organisations by correlating disparate fragments of information from many sources [41].

Consequently, Big Data analytics solutions must be engineered to minimise their effects on data reutilization and privacy-preserving policies. Furthermore, there exists a further concern with data security, namely Provenance. Provenance is defined in the literature as the origin of data. This is seen as a significant difficulty for Big Data analytics systems overall, and specifically for cyber-security analytics [42]. While Big Data enhances the diversification of data sources for processing, it remains questionable if these sources satisfy the reliability standards necessary for the analytics algorithms used by the suggested analytical solutions to provide accurate and trustworthy outcomes. Generally, authenticity and integrity standards must be enforced on data from various sources prior to its utilisation in analytics. The problem is exacerbated when harmful data is introduced

into the Big Data pool, hindering the extraction of insights essential for making significant and key choices [43].

3. Application examples

A healthcare organization collects and analyzes massive amounts of patient data, including medical records, diagnostic reports, and treatment histories, stored in a big data platform. The organization needs to ensure the security of this sensitive data while complying with regulations like HIPAA (Health Insurance Portability and Accountability Act).

Proactive Security Model Implementation:

Threat Intelligence and Predictive Analytics:

- Use machine learning algorithms to analyze historical data breaches and identify patterns that could indicate potential threats.
- Deploy predictive analytics to forecast possible attack vectors, such as ransomware targeting patient records or insider threats.

Real-Time Monitoring and Anomaly Detection:

- Implement real-time monitoring tools to track data access and usage patterns.
- Use anomaly detection systems to flag unusual activities, such as unauthorized access attempts or large data transfers.

Data Encryption and Tokenization:

- Encrypt sensitive patient data both at rest and in transit to prevent unauthorized access.
- Use tokenization to replace sensitive data with non-sensitive equivalents, reducing the risk of exposure during data processing.

Access Control and Authentication:

- Implement role-based access control (RBAC) to ensure only authorized personnel can access specific data.
- Use multi-factor authentication (MFA) to add an extra layer of security for user logins.

Automated Patch Management:

- Regularly update and patch software and systems to address known vulnerabilities.
- Use automated tools to ensure timely application of security patches across the big data infrastructure.

User Training and Awareness:

- Train employees on cybersecurity best practices and the importance of protecting patient data.
- Conduct regular phishing simulations to educate staff on recognizing and avoiding social engineering attacks.

4. Comparative Analysis Of AI-Based Big Data Security Models.

A comparative analysis of AI-based big data security models involves evaluating different approaches, techniques, and frameworks that leverage artificial intelligence (AI) to enhance the security of big data systems. Such an analysis typically compares the strengths, weaknesses, and applicability of various models in addressing security challenges like data breaches, unauthorized access, and cyberattacks. Below is a structured review of such a comparative analysis:

Key Areas of Comparison in AI-Based Big Data Security Models

1. **Techniques and Algorithms Used:**
   - Machine Learning (ML): Models like decision trees, random forests, and support vector machines (SVMs) are commonly used for anomaly detection and classification of threats.
   - Deep Learning (DL): Neural networks, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are effective for complex pattern recognition and real-time threat detection.
   - Natural Language Processing (NLP): Used for analyzing unstructured data, such as logs or emails, to detect phishing or social engineering attacks.

- Reinforcement Learning (RL): Applied in dynamic environments where the system learns to respond to threats through trial and error.

2. **Use Cases and Applications:**

- Anomaly Detection: Identifying deviations from normal behavior in large datasets.
- Predictive Analytics: Forecasting potential security threats based on historical data.
- Fraud Detection: Detecting fraudulent activities in financial transactions or user behavior.
- Intrusion Detection: Monitoring network traffic for signs of cyberattacks.

3. **Performance Metrics:**

- Accuracy: How well the model identifies true threats and avoids false positives.
- Scalability: The ability to handle large volumes of data efficiently.
- Latency: The time taken to detect and respond to threats.
- Resource Efficiency: The computational and memory requirements of the model

4. **Comparative Analysis Table:**

Table (1): Comparative Analysis among the methods and techniques of modelling big data.

Model	Technique	Strengths	Weaknesses	Best Use Case
Machine Learning	Decision Trees, SVM	Interpretable, efficient for structured data	Limited scalability, struggles with unstrucured data	Fraud detection, anomaly detection
Deep Learning	CNNs, RNNs	High accuracy, handles complex patterns	Computationally expensive, requires large datasets	Intrusion detection, predictive analytics
Hybrid Models	ML + DL	Combines strengths of both approaches	Increased complexity, resource-intensive	Comprehensive threat detection
Reinforcement Learning	Q-learning, Deep Q-networks	Adapts to dynamic environments	Requires extensive training, high latency	Real-time threat response

5. **The Impact of Potential False Positives/Negatives in Predictive AI**

The impact of false positives and false negatives in predictive AI systems can be significant, depending on the application domain. These errors arise when the model makes incorrect predictions:

- False Positive (Type I Error): The model predicts a positive outcome when the actual outcome is negative.
- False Negative (Type II Error): The model predicts a negative outcome when the actual outcome is positive.

The consequences of these errors vary widely based on the context in which the AI system is deployed. Below is a detailed analysis of their impact across different domains:

1. Healthcare

In healthcare, predictive AI is used for disease diagnosis, patient monitoring, and treatment recommendations.

A. False Positives: Impact: Unnecessary medical interventions, such as surgeries, medications, or additional tests, which can lead to: Increased healthcare costs. Physical and emotional stress for patients. Overburdening of medical resources. Example: A false positive in cancer screening could lead to unnecessary biopsies or chemotherapy.

B. False Negatives: Impact: Missed diagnoses, which can result in: Delayed treatment, worsening the patient's condition. Increased mortality rates. Legal and ethical consequences for healthcare providers. Example: A false negative in COVID-19 testing could lead to infected individuals spreading the virus.

2. Finance

In finance, AI is used for fraud detection, credit scoring, and investment predictions.

- False Positives: Impact: Legitimate transactions flagged as fraudulent, leading to: Customer dissatisfaction and loss of trust. Blocked transactions, causing inconvenience. Increased operational costs for manual review. Example: A false positive in fraud detection could block a legitimate credit card transaction.

3. Cybersecurity

AI is used to detect cyber threats, such as malware, phishing, and network intrusions.

A. False Positives: Impact: Legitimate activities flagged as threats, causing: Disruption of normal operations. Wasted time and resources investigating non-threats. Reduced trust in the security system. Example: A false positive in intrusion detection could block legitimate users from accessing a network.

B. False Negatives: Impact: Actual threats going undetected, leading to: Data breaches and loss of sensitive information. Financial and reputational damage. Legal consequences for failing to protect data. Example: A false negative in malware detection could allow a ransomware attack to succeed.

4. Criminal Justice

AI is used for predictive policing, risk assessment, and parole decisions.

A. False Positives: Impact: Innocent individuals flagged as high-risk, resulting in: Unjust surveillance or arrests. Erosion of trust in law enforcement. Social and psychological harm to individuals. Example: A false positive in predictive policing could lead to wrongful targeting of individuals.

B. False Negatives: Impact: High-risk individuals not being identified, leading to: Increased crime rates. Public safety concerns. Criticism of the justice system's effectiveness. Example: A false negative in risk assessment could release a dangerous offender on parole.

5. Marketing and Customer Engagement

AI is used for customer segmentation, churn prediction, and personalized recommendations.

A. False Positives: Impact: Incorrectly identifying customers as likely to churn or interested in a product, leading to: Wasted marketing resources. Customer annoyance from irrelevant offers. Reduced effectiveness of campaigns. Example: A false positive in churn prediction

could result in unnecessary discounts to retain loyal customers.

B. False Negatives: Impact: Failing to identify customers who are likely to churn or interested in a product, resulting in: Lost revenue opportunities. Decreased customer satisfaction. Reduced competitive advantage. Example: A false negative in recommendation systems could miss recommending a product a customer would have purchased.

Conclusion

Cyber-attacks are becoming more complex owing to fast technological breakthroughs, complicating the timely and effective mitigation of these risks. Conventional cyber-security systems, including log management, traditional IDS, IPS, and SIEM tools, are unable to address emerging threats and techniques. Moreover, the emergence of Big Data, characterised by its vast volume, rapid creation rate, and many data kinds, renders standard technologies inadequate for detecting cyber risks in this setting. Consequently, Big Data analytics solutions are essential for mitigating such complex hazards in Big Data. To be effective, Big Data analytics cybersecurity solutions must meet some fundamental criteria. They must handle data originating from diverse sources and implement superior management solutions for large-scale data. They must also manage various data kinds and effectively visualise it to provide quick and straightforward conclusions and insights extraction. To meet these criteria, Big Data analytics cybersecurity solutions must include a robust and high-performance architecture that enables the management of Big Data across many contexts. Several Big Data analytics safety measures have been developed and implemented by major industrial corporations, including IBM and Teradata. Nevertheless, several problems hinder the comprehensive implementation of Big Data analytics for cybersecurity. These issues include the complexity of managing unstructured and intricate data, as well as the need for real-time and streaming data analysis. The adoption is further impeded by concerns over data privacy and provenance, as well as the need for adaptation to dynamic changes in data behaviours. To promote its adoption, future initiatives include improving privacy and security measures that conceal important security-related information. Furthermore, behavioural analytics have to be integrated with Big Data analytics to effectively mitigate insider risks. Additional visualisation tools are essential for security analysts to get valuable early insights on cyber threats and security breaches for subsequent inquiry. The effective adoption must take into account the relevance to diverse IoT applications.

References

[1] D. Jeon, “Analysis model for prediction of cyber threats by utilizing Big Data Technology,” The Journal of Korean Institute of Information Technology, vol. 12, no. 5, May 2014. doi:10.14801/kiitr.2014.12.5.81

[2] A. Aigner and A. Khelil, “A semantic security model for cyber-physical systems to identify and evaluate potential threats and vulnerabilities,” Proceedings of the 7th International Conference on Internet of Things, Big Data and Security, pp. 249–257, 2022. doi:10.5220/0011086300003194

[3] R. Reddy Palle, “Explore the application of predictive analytics and machine learning algorithms in identifying and preventing cyber threats and vulnerabilities within Computer Systems,” International Journal of Science and Research (IJSR), vol. 12, no. 2, pp. 1704–1712, Feb. 2023. doi:10.21275/es24101104007

[4] A. Shalaginov and T.-M. Gronli, “Securing smart future: Cyber threats and intelligent

means to respond," 2021 IEEE International Conference on Big Data (Big Data), Dec. 2021. doi:10.1109/bigdata52589.2021.9671703

[5] M. Krishna Pasupuleti, "Ai and Big Data for Climate Resilience: Predictive Analytics in environmental management," AI and Big Data in Climate Change: Predictive Analytics for Environmental Management, pp. 255–280, Nov. 2024. doi:10.62311/nesx/46601

[6] H. Mohamed, A. El Bolock, and C. Sabty, "IntrusionHunter: Detection of cyber threats in Big Data," Proceedings of the 12th International Conference on Data Science, Technology and Applications, pp. 311–318, 2023. doi:10.5220/0012081900003541

[7] Y. G. Zeng, "Identifying email threats using predictive analysis," 2017 International Conference on Cyber Security And Protection Of Digital Services (Cyber Security), Jun. 2017. doi:10.1109/cybersecpods.2017.8074848

[8] R. Wang, "Ai-powered predictive cybersecurity in identifying emerging threats through machine learning," 2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), vol. 14, pp. 819–825, Feb. 2024. doi:10.1109/eebda60612.2024.10485789

[9] I. Oncioiu, A. G. Petrescu, D. A. Mândricel, and A. M. Ifrim, "Proactive information security strategy for a secure business environment," Advances in Data Mining and Database Management, pp. 214–231, 2019. doi:10.4018/978-1-5225-7277-0.ch012

[10] T. Hirashima, A. A. Supianto, and Y. Hayashi, "Model-based approach for educational big data analysis of learners thinking with Process Data," 2017 International Workshop on Big Data and Information Security (IWBIS), pp. 11–16, Sep. 2017. doi:10.1109/iwbis.2017.8275096

[11] M. Fereidouni, A. Mosharrof, U. Farooq, and A. B. Siddique, "Proactive prioritization of APP issues via Contrastive Learning," 2022 IEEE International Conference on Big Data (Big Data), vol. 32, pp. 535–544, Dec. 2022. doi:10.1109/bigdata55660.2022.10020586

[12] D. E. Holmes, "7. Big Data Security and the snowden case," Big Data: A Very Short Introduction, pp. 90–104, Nov. 2017. doi:10.1093/actrade/9780198779575.003.0007

[13] Gousiya Begum, S. Z. Huq, and A. P. Kumar, "Sandbox security model for Hadoop File System," Journal of Big Data, vol. 7, no. 1, Sep. 2020. doi:10.1186/s40537-020-00356-z

[14] F. Alshanik, A. Apon, Y. Du, A. Herzog, and I. Safro, "Proactive query expansion for streaming data using external sources," 2022 IEEE International Conference on Big Data (Big Data), vol. 85, pp. 701–708, Dec. 2022. doi:10.1109/bigdata55660.2022.10020577

[15] S. Singh and D. Kumar, "Data Fortress: Innovations in big data analytics for proactive cybersecurity defense and asset protection," International Journal of Research Publication and Reviews, vol. 5, no. 6, pp. 1026–1031, Jun. 2024. doi:10.55248/gengpi.5.0624.1425

[16] S. Singh and D. Kumar, "Data Fortress: Innovations in big data analytics for proactive cybersecurity defense and asset protection," International Journal of Research Publication and Reviews, vol. 5, no. 6, pp. 1026–1031, Jun. 2024. doi:10.55248/gengpi.5.0624.1425

[17] B. Yang and T. Zhang, "A scalable meta-model for Big Data Security analyses," 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High

Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), vol. 9137, pp. 55–60, Apr. 2016. doi:10.1109/bigdatasecurity-hpsc-ids.2016.71

[18] S. R. Islam, S. Khaled Ghafoor, and W. Eberle, “Mining illegal insider trading of stocks: A proactive approach,” 2018 IEEE International Conference on Big Data (Big Data), Dec. 2018. doi:10.1109/bigdata.2018.8622303

[19] L. Yue, H. Junqin, Q. Shengzhi, and W. Ruijin, “Big data model of security sharing based on Blockchain,” 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM), Aug. 2017. doi:10.1109/bigcom.2017.31

[20] D. Martínez-Mosquera and S. Luján-Mora, “Data cleaning technique for Security Big Data Ecosystem,” Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security, pp. 380–385, 2017. doi:10.5220/0006360603800385

[21] E. Bertino, “Big Data Security and Privacy,” 2016 IEEE International Conference on Big Data (Big Data), Dec. 2016. doi:10.1109/bigdata.2016.7840581

[22] P. Neelakrishnan, “Future ready data security,” Autonomous Data Security, pp. 329–354, 2024. doi:10.1007/979-8-8688-0838-8_7

[23] S. Koley, “6. big data security issues with challenges and solutions,” Big Data Security, pp. 95–142, Oct. 2019. doi:10.1515/9783110606058-006

[24] S. Koley, “6. big data security issues with challenges and solutions,” Big Data Security, pp. 95–142, Oct. 2019. doi:10.1515/9783110606058-006

[25] P. Neelakrishnan, “Traditional Data Security,” Autonomous Data Security, pp. 41–86, 2024. doi:10.1007/979-8-8688-0838-8_2

[26] S. Munirathinam and B. Ramadoss, “Big data predictive analtyics for proactive semiconductor equipment maintenance,” 2014 IEEE International Conference on Big Data (Big Data), vol. 30, pp. 893–902, Oct. 2014. doi:10.1109/bigdata.2014.7004320

[27] M. Chen, “A hierarchical security model for multimedia Big Data,” Big Data, pp. 441–453, 2016. doi:10.4018/978-1-4666-9840-6.ch022

[24] J. F. DeFranco and B. Maley, “Cyber Security and Digital Forensics Careers,” *What Every Engineer Should Know About Cyber Security and Digital Forensics*, pp. 135–154, Sep. 2022. doi:10.1201/9781003245223-9

[25] S. Rudd, “Ransomware reconnaissance: Interrogating certificates towards proactive threat mitigation,” *Proceedings of the 9th International Conference on Internet of Things, Big Data and Security*, pp. 97–106, 2024. doi:10.5220/0012710600003705

[26] S. Munirathinam and B. Ramadoss, “Big data predictive analtyics for proactive semiconductor equipment maintenance,” *2014 IEEE International Conference on Big Data (Big Data)*, vol. 30, pp. 893–902, Oct. 2014. doi:10.1109/bigdata.2014.7004320

[27] S. Koley, “6. big data security issues with challenges and solutions,” *Big Data Security*, pp. 95–142, Oct. 2019. doi:10.1515/9783110606058-006

[28] E. Bertino, “Big Data Security and Privacy,” 2016 IEEE International Conference on Big Data (Big Data), Dec. 2016. doi:10.1109/bigdata.2016.7840581

[29] D. Martínez-Mosquera and S. Luján-Mora, “Data cleaning technique for Security Big Data Ecosystem,” *Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security*, pp. 380–385, 2017. doi:10.5220/0006360603800385

[30] B. Yang and T. Zhang, “A scalable meta-model for Big Data Security analyses,” *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, vol. 9137, pp. 55–60, Apr. 2016. doi:10.1109/bigdatasecurity-hpsc-ids.2016.71

[31] Mohan, Pulkit, et al. "Cybersecurity and Nuclear Facilities." The Challenges of Nuclear Security (2024): 245.

[32] M. Saeed and N. Arshed, “Big Data Analytics adoption in the Indian insurance industry: Challenges and Solutions,” *Big Data Analytics in the Insurance Market*, pp. 81–102, Jul. 2022. doi:10.1108/978-1-80262-637-720221005

[33] B. Pokorny, “Cybersecurity training,” Big Data Analytics in Cybersecurity, pp. 115–136, Sep. 2017. doi:10.1201/9781315154374-6

[34] L. Harrison, “Data Visualization for cybersecurity,” Big Data Analytics in Cybersecurity, pp. 99–114, Sep. 2017. doi:10.1201/9781315154374-5

[35] B. Pokorny, “Cybersecurity training,” Big Data Analytics in Cybersecurity, pp. 115–136, Sep. 2017. doi:10.1201/9781315154374-6

[36] D. Caragea and X. Ou, “Big Data Analytics for Mobile app security,” Big Data Analytics in Cybersecurity, pp. 169–184, Sep. 2017. doi:10.1201/9781315154374-8

[37] S. Luo, M. B. Salem, and Y. Zhai, “The power of Big Data in cybersecurity,” Big Data Analytics in Cybersecurity, pp. 3–22, Sep. 2017. doi:10.1201/9781315154374-1

[38] J. Deng and O. Savas, “Data and Research Initiatives for cybersecurity analysis,” *Big Data Analytics in Cybersecurity*, pp. 309–328, Sep. 2017. doi:10.1201/9781315154374-14

[39] S. K. Gupta, O. Hrybiuk, N. S. Cherukupalli, and A. K. Shukla, “Big data analytics tools, challenges and its applications,” Smart Cities, pp. 307–320, Oct. 2023. doi:10.1201/9781003376064-16

[40] W. Han and Y. Xiao, “Cybersecurity in internet of things (IOT),” Big Data Analytics in Cybersecurity, pp. 221–244, Sep. 2017. doi:10.1201/9781315154374-10

[41] M. Balmakhtar and S. E. Mensch, “Big data analytics adoption factors in improving information systems security,” Research Anthology on Big Data Analytics, Architectures, and Applications, pp. 1231–1248, 2022. doi:10.4018/978-1-6684-3662-2.ch059

[42] S. Aggarwal and S. Sindakis, “Big Data Analytics and cybersecurity: Emerging trends,” Big Data Analytics in Cognitive social media and Literary Texts, pp. 151–164, 2021. doi:10.1007/978-981-16-4729-1_8.

[43] K. Kaushik, “Blockchain enabled Artificial Intelligence for cybersecurity systems,” Studies in Big Data, pp. 165–179, 2022. doi:10.1007/978-3-031-05752-6_11