Scientific Publications

Journal of Artificial Intelligence and Big Data, 2022, 2, 1125
www.scipublications.org/journal/index.php/jaibd
DOI: 10.31586/jaibd.2022.1125

Article

Big Data and AI/ML in Threat Detection: A New Era of Cybersecurity

Seshagirirao Lekkala [1,*], Raghavaiah Avula [2], Priyanka Gurijala [3]

[1] Sr Software Engineer, USA
[2] Sr. Principal Engineer, Palo Alto Networks Inc, USA
[3] Software Engineer, Microsoft Corporation, USA

*Correspondence: Seshagirirao Lekkala (seshagiriraolekkala@outlook.com)

Abstract: The unrelenting proliferation of data, entwined with the prevalence of mobile devices, has given birth to an unprecedented growth of information obscured by noise. With the Internet of Things and myriad endpoint devices generating vast volumes of sensitive and critical data, organizations are tasked with extracting actionable intelligence from this deluge. Governments and enterprises alike, even under pressure from regulatory boards, have strived to harness the power of data and leverage it to enhance safety and security, maximize performance, and mitigate risks. However, the adversaries themselves have capitalized on the unequal battle of big data and artificial intelligence to inflict widespread chaos. Therefore, the demand for big data analytics and AI/ML for high-fidelity intelligence, surveillance, and reconnaissance is at its highest. Today, in the cybersecurity realm, the detection of adverse incidents poses substantial challenges due to the sheer variety, volume, and velocity of deep packet inspection data. State-of-the-art detection techniques have fallen short of detecting the latest attacks after a big data breach incident. On the other hand, computational intelligence techniques such as machine learning have reignited the search for solutions for diverse monitoring problems. Recent advancements in AI/ML frameworks have the potential to analyze IoT/edge-generated big data in near real-time and assist risk assessment and mitigation through automated threat detection and modeling in the big data and AI/ML domain. Industry best practices and case studies are examined that endeavor to showcase how big data coupled with AI/ML unlocks new dimensions and capabilities in improved vigilance and monitoring, prediction of adverse incidents, intelligent modeling, and future uncertainty quantification by data resampling correction. All of these avenues lead to enhanced robustness, security, safety, and performance of industrial processes, computing, and infrastructures. A view of the future and how the potential threats due to the misuse of new technologies from bandwidth to IoT/edge, blockchain, AI, quantum, and autonomous fields is discussed. Cybersecurity is again playing out at a pace set by adversaries with low entry barriers and debilitating tools. The need for innovative solutions for defense from the emerging threat landscape, harnessing the power of new technologies and collaboration, is emphasized.

Keywords: Big Data, AI/ML (Artificial Intelligence/Machine Learning), Cybersecurity, IoT (Internet of Things), Threat Detection, Data Analytics, Deep Packet Inspection, Risk Mitigation, Automated Threat Detection, Computational Intelligence

How to cite this paper: Lekkala, S., Avula, R., & Gurijala, P. (2022). Big Data and AI/ML in Threat Detection: A New Era of Cybersecurity. *Journal of Artificial Intelligence and Big Data, 2*(1), 32–48. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1125

Received: July 15, 2022
Revised: September 16, 2022
Accepted: October 12, 2022
Published: October 15, 2022

1. Introduction

The proliferation of cyberattacks in frequency, sophistication, and impact has raised serious concerns about the vulnerabilities of an ever-growing dependency on cyberspace. Cybercrime has been recognized as a major threat that needs to be addressed as a global priority, similar to perennial concerns such as terrorism and organized crime. There is a stark rise in cyber threat capability and willingness among the cybercrime community. Complaints about cyberattacks have significantly increased compared to the prior year.

Business Email Compromise was one example of an emerging threat that caused reputable organizations to lose hundreds of millions of dollars to cybercrime syndicates. The duration and effects of the pandemic have also brought deeper potential issues, such as stress-induced impulsive behavior due to sudden reorganizations of business models and work-from-home practices.

Network-based attacks, from cryptographic malware to rapid and automated activities, are becoming easier due to an increase in crime-as-a-service. Sophisticated zero-day attacks and new data exfiltration techniques are emerging with the phrase Ransomware-as-a-Service. The traditional update-and-protect stance is becoming impractical with the advance of attack technologies, such as state-of-the-art machine learning techniques in evasion and similar behavioral obfuscations. Threat intelligence has to be carried and shared globally to face leading, spearheaded offensives. Every cyber defensive upgrade cannot be just reactive patchwork since bought technologies cannot keep up with fast-moving attacks. The need for building an adaptable cyber defense has been realized.

Monitoring anomalies in cyber environment behaviors currently stands as the de facto approach for building a new defense. A paradigm shift from a close-inside trust model to an outside view and building a defense based on observing anomalies at all possible behavior monitoring points across attack purposes, vehicle infrastructures, and user activities is envisioned as a foundation for new defensive systems. Automated and intelligent deep machine learning technologies must be introduced to analyze such intractable volumes of diverse data and threats. This work explores how big data and AI machine learning can be utilized to revolutionize anomaly detection in threat detection and bring previously unthinkable corporations to understand profile baselines and detect emerging threats.

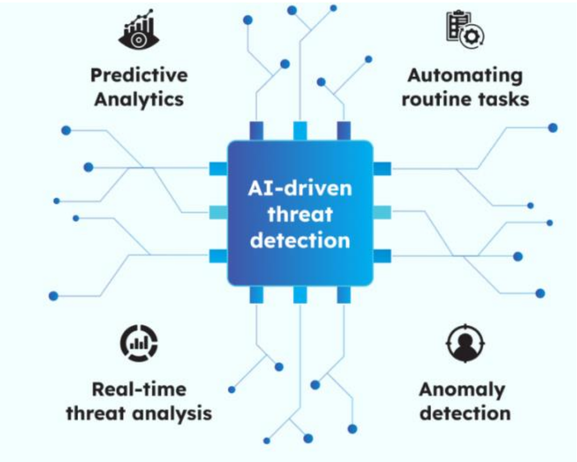


Figure 1. Threat Detection

1.1. Background and Significance

With the evolution of Internet-enabled service infrastructures, the massive increase in online business activities, the proliferation of smart devices, and IoT adoption, the world is witnessing tremendous amounts of data being generated constantly. Experts estimate that by 2025, global daily data generation will surpass 463 exabytes, with 40 trillion gigabytes of data stored. By 2030, it is reported that IoT will account for 500 billion internet-connected devices, feeding 79.4 ZB, or 80 trillion gigabytes, of ads. Handling, processing, saving, and storing these huge troves of data continues to pose serious challenges to database and computer network researchers. Such challenges include but are not limited to how to efficiently process massive volumes of multi-modal data in motion, how to extract actionable knowledge from massive data flows at scale, and how to provide high-level and privacy-compliant interfaces for the general public to gain access to big data [5].

As the data considered "big" becomes ubiquitous, sensors, cameras, and smart devices are being assimilated into every aspect of daily life. Such fast and exhaustive data generation comes with a dark side, as many data sources, environment scanning devices, and city infrastructures become targets for malicious and illegal activities, leading to increasing threats to the lifestyle of a nation, a company, or an individual. The aftermath of a big data breach can take a heavy toll on organizations and their customers. In just a couple of weeks, a breach can damage the organization's reputation, stock price, brand equity, and market competitiveness. As such, cyber-attacks, data breaches, and anomalies represent a significant threat to the economy and societal security. In, the increasing sophistication of threats, capability enhancement of threat actors, and a significant data breach used to monitor, infiltrate, and locate targets are revealed. Such incidents emphasize that conventional firewall-based systems are no longer sufficient to combat dynamically evolving threats. The key to protecting from such threats is to gain awareness of the potential threats to the data itself and to deploy mechanisms to detect and neutralize caused actions or events. Even though many modern organizations have invested heavily in IT infrastructures like firewalls, proxy servers, and VPNs, attacks on electronic data and systems still successfully occur. These tools can manage commonplace network-based intrusions but are unable to address sophisticated data breaches [7].

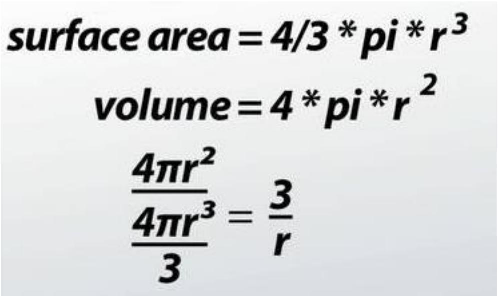
1.2. Research Objectives and Scope

In recent times, cyberspace has emerged as a new domain of conflict. An integral part of cyberspace is the Internet, which is a ubiquitous part of modern life. As the world becomes more interconnected through technology, cybersecurity has become a growing concern as the cyber domain becomes a state-of-the-art battlefield. Over the last decade, cyberattacks have increased massively in number, sophistication, and severity. Data breaches, insider threats, malware, DoS attacks, ransomware, and inappropriate data accessibility have become part of everyday enterprise life. These vulnerabilities lead to the loss of stakeholders' trust, reputation damage, financial losses, and possible legal actions for the victimized entity. To deal with increasing cyberattacks and improve a network's security posture, organizations have started employing AI and ML algorithms for their threat and attack detection. Machine learning with big data assists organizations in the semi-automated analysis of security alerts and monitoring of hazardous activities over the network [1].

The objective of this thesis is to focus on the incorporation of big data and AI/ML techniques in cybersecurity to ensure a sustainable network security posture. Thus, an investigation of various machine learning methods for threat detection over big data cybersecurity datasets would be conducted, and the best approaches among them would be uncovered, which can then be utilized for alarm analysis in comprehensive threat detection systems. With growing complexities in peril and attack definitions, data stratifications can help enhance network security. Thus, data categorization methodologies will also be investigated in intrusion detection. Notably, this thesis will try to uncover how both penetration and zero-day attacks can be identified in a network. This comprehensive investigation is aimed at assisting growing enterprises in setting up a robust cybersecurity framework and is expected to bring a new era of network security worldwide [2].

The scope of the research is confined to the inclusion of AI/ML and big data in cybersecurity exclusively towards threat and attack detection. Additionally, only networks running in the TCP/IP layer of the OSI model would be explored. Lastly, the analysis of potential strategies and techniques will be conducted on open-source datasets that are realizable in commonly used programming languages.

Equation 1: Surface Area to Volume Ratio



$$surface\ area = 4/3 * pi * r^3$$

$$volume = 4 * pi * r^2$$

$$\frac{4\pi r^2}{\frac{4\pi r^3}{3}} = \frac{3}{r}$$

2. Foundations of Big Data and AI/ML in Cybersecurity

The tremendous volume of information presently in existence has spurred a comprehensive movement towards the collecting and analysis of data, powered by accelerated advances in computing technologies, data management infrastructures, data analytics algorithms, and visualization methods. Such forces combine to enable the processing and understanding of vast amounts of numerical, textual, and sensory data, streaming continuously and in varying forms from all walks of life on a 24/7 basis. The insight gained from big data can augment both organization-wide and social processes, informing business operations and public policy, thereby rendering society more knowledge- and information-centered [9].

Academically, growing interest in big data has occurred concurrently with increased attention on the notion of social media, which simply denotes a range of broadly viewed social networking platforms. Investigating big data and social media, particularly concerning their intersection and potential synergy, is of interest to both business practitioners and academics. The overarching concept of big data and social media pertains to an extensive and multi-faceted exploration of how large amounts of data produced by expansive online social networking platforms can be analyzed to add value to organizations, corporations, and society [3].

The vast majority of current information is in an unstructured or semi-structured format, such as emails, videos, and social network messages. A big data management problem revolves around defining the right big data-gaining paradigm and format transformation mechanism by which structured data can be obtained to run standard analytical models. Big data can be defined as data that becomes too complex to manage, analyze, and visualize with common database management tools and analytical models. Big data is characterized by the so-called three Vs—volume, velocity, and variety.

Artificial intelligence (AI) is intelligence exhibited by machines, in contrast to the natural intelligence displayed by humans and animals. Leading AI textbooks define the field as the study of intelligent agents: any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. Colloquially, the term artificial intelligence is often used to describe machines that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving. As machines become increasingly capable, tasks considered to require intelligence are often removed from the definition of AI, a phenomenon known as the AI effect [4].

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being programmed to do so. The core objective is to allow computers to learn automatically without human intervention or assistance. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. Machine learning algorithms use computational methods to learn information directly from data without relying on a predetermined equation as a model. Instead, they rely on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention [6].

2.1. Conceptual Overview of Big Data

As we enter the new era of big data, organizations face challenges to fully exploit the vast amount of data to build a competitive edge and ensure the long-term sustainability of business. There are significant technical challenges, mostly coming from the 6 Vs of big data, which include: Volume – Refers to large amounts of data that can overwhelm traditional database systems with relatively lower records or observations; Velocity – Data is continuously generated at high rates and needs to be processed in real-time before it becomes useless; Variety – Different forms of data coming from various sources in different formats; Variability – Fluctuations in data streams and targets which are inconsistent and can hinder expectations; Veracity – Accuracy, precision, cleansed, and other quality metrics that need to be considered for proper analysis; and Value – Need for data enrichment, data fusion, and aggregation to improve data quality [10].

Although there are several commercial databases available, some core challenges need to be considered while deciding on the NoSQL solution. First, the CAP theorem highlights that a consistent state of data cannot be guaranteed under a partitioned network, and hence only two other aspects are guaranteed to be considered: either having a linear consistent state achievable under normal operations only or using mechanisms that allow nodes to readily diverge on what the consistent state is through high-frequency updates, eventual consistency, etc. The fault tolerance model of the database is crucial to avoid data loss and maintain consistency in operation in case of failure. Queries need to be well understood since complex queries can reduce performance considerably in some NoSQL systems such as key-value stores. The hardware and partitioning strategy need to be decided as well in shared or structured partitions. Security and data geo-distribution also need to be considered either in the first choice of database or in the organization remotely supporting a database appliance as a service provider. Finally, operational concerns are needed on how to back up and restore a database in a consistent manner or how to identify potential hotspot scenarios affecting system performance and even availability [8].

With the advent of the internet and other technologies, opportunities for collecting data have risen exponentially. From web logs, social media feeds, point-of-sale terminals, biosensors, RFID cards, and geographic position from GPS sensors, companies have started to collect and process huge amounts of data, trying to make sense of them in a business intelligence fashion. Despite processing all of this data, companies are losing a lot of value if proper measures are not taken to detect these entities and meaningfully represent them. Realizing this extraction of valuable knowledge from data is possible through properly curating a database. As data is not a single database point but rather persistent knowledge of an addressable object, the same model is used to detect entities in a big data context [15].

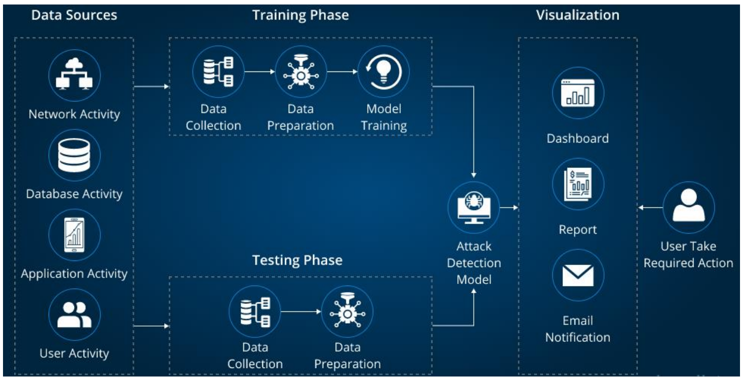


Figure 2. An overview of Data security in AI systems

2.2. Fundamentals of Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) is a technology that enables machines to mimic human intelligence, including cognitive functions such as learning, reasoning, problem-solving, perception, language understanding, and interaction. AI can be categorized into two types: narrow (or weak) AI and general (or strong) AI. Narrow AI systems, such as smart speakers and self-driving cars, emulate human performance in a specific domain but do not possess general intelligence. General AI systems, still theoretical, would emulate human cognitive performance in multiple domains. Machine Learning (ML) is a subfield of AI focused on developing algorithms that allow machines to learn from data and improve performance over time. ML systems can identify patterns, make decisions, and predict outcomes without explicitly programmed instructions. ML can be categorized into three types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning algorithms learn from labeled datasets, making predictions based on input features. Unsupervised learning algorithms analyze unlabeled datasets without predefined categories or output variables, identifying clusters or relations between features. Reinforcement learning is a trial-and-error approach where agents learn by interacting with an environment and receiving rewards or penalties based on their actions[11].

Deep Learning (DL) is a subfield of machine learning that uses neural networks with multiple layers of processing nodes to learn complex representations of data. It is applied in computer vision, image recognition, speech recognition, natural language processing, and game playing. In the context of cybersecurity, AI and machine learning solutions can be broadly classified into three types: (1) solutions that apply AI and machine learning analytic techniques to improve the performance of traditional detection techniques, (2) solutions that incorporate AI and machine learning analytic techniques into the detection techniques, and (3) solutions that tackle cybersecurity problems through AI and machine learning analytic techniques. The first type focuses on improving the performance of well-known methods such as misbehavior detection through clustering techniques and more robust anomaly detection models using Bayesian statistics. AI and machine learning analytic approaches typically improve the efficiency of techniques based on a modeling foundation. The second type encompasses solutions that do not have a known modeling foundation, such as honeypots or IP blacklists. AI and machine learning analytic solutions in this category can be applied to gather cleaner, more accurate data efficiently and possibly from other sources to enhance the robustness of the detection technique. The third type involves tackling problems in cybersecurity areas with suitable AI and machine learning techniques, such as bot detection with support vector machines or spam filtering with natural language processing [13].

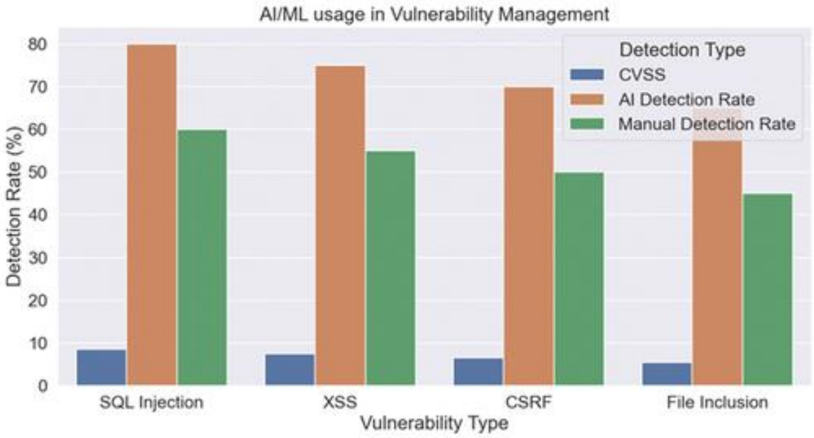


Figure 3. Current trends in AI and ML for cybersecurity

3. The Role of Big Data in Cybersecurity Threat Detection

Current cybersecurity systems find it more difficult to corral the growing surfeit of data stemming from networks, devices, and users. Network devices and applications generate massive amounts of varied data. Businesses are constantly communicating and exchanging data with other organizations. Data comes from sources as disparate as open government databases, publicly available social media interactions, beta testers on the lookout for zero-day vulnerabilities, and security researchers performing assurance audits on commonly used software. All of this data must be processed, analyzed, and correlated in real time if organizations are to defend themselves from cyber threats. Big Data technologies provide the analytics needed to connect the dots between seemingly unrelated events [16].

Before deploying the big data framework, the organization must design its key components and infrastructure. This includes designing the collection and storage architecture, data preprocessing pipelines, the indexing scheme, analytics to develop knowledge, action or orchestration pipelines, vulnerability and incident correlation processes, reports on SMEs and management dashboards, knowledge, alerts, and incident sharing architectures, and integration with current monitoring solutions. Insight into all major components will mitigate the time-consuming trial and error of both implementing and adjusting the components of a big data framework for threat detection.

Data collection entails gathering threat detection candidate data and placing it in a big data storage framework. To prevent deploying costly technologies to store unstructured data alone, proper tuning of data collection should take into account the organization's threat landscape, tangible technology assets, and policies. In deploying a data collection architecture, the proponent must determine which internal and external data streams to ingest and store, social media and vendor advisories, as well as incident reports, and which sensing technologies to deploy, host firewalls, network packet capture, or deep packet inspection technology [19].

For data collection and preprocessing to be truly effective, organizations must clearly understand the data sources and data types being collected. After collecting relevant raw data feeds, it must be transformed into a clear structure for later use. At this stage, various formats must be paraphrased to text, special characters erased, timestamps put into the same format, and so forth. Data type specifications include the big traditional types and custom data types. To augment event structure, categorical data must be encoded as one-hot encodings, periodically updated data must have time-based encodings added, and geolocated data should include encoded freight movements [22].

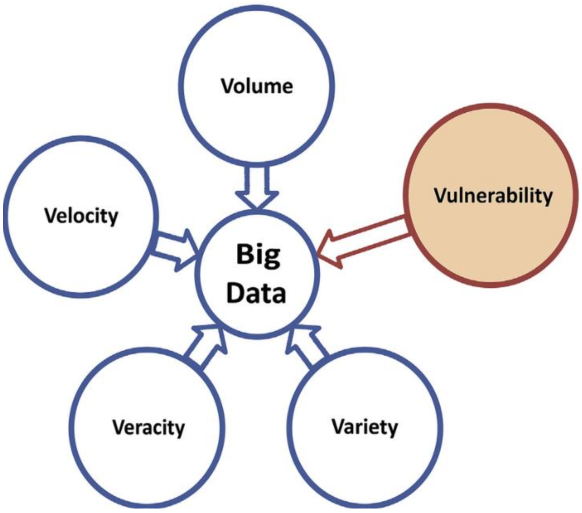


Figure 4. Cybersecurity in Big Data Era

3.1. Data Collection and Storage

Data collection is the process of gathering relevant data from an organization's IT systems for analysis. It can be done on a local system or via the cloud and can often be automated. Data is typically stored in a staging area to wait for further processing.

Data sources for cybersecurity generally come from both hardware and software within the organization's IT infrastructure. This typically includes network sensors, security devices, user devices, firewalls, proxies, intrusion detection and prevention systems, antivirus software, email gateways, web gateways, domain controllers, DHCP servers, etc. These data sources are monitored to report any anomalous or malicious activity. In addition to monitoring data sources, contextual data that provides information about the organization, data sources, users, assets, and classes of monitoring data can be used for analysis. Event data from outside networks can also be used to supplement the monitoring of the targeted organization's IT infrastructure.

Depending on the data source, reporting can be continuous or at intervals. The reporting data can be raw logging data, summary statistics, alerts, or combinations thereof. In many cases, it can also be difficult to extract data from remote systems due to network restrictions, data formats, and organizational policies. Some organizations may employ a data warehouse that collects and stores important compliance-related data for regulatory purposes [18].

Various types of data can be collected from network sensors. Packet captures represent the full unmodified network protocol data. NetFlow accounting data represent summary statistics about connections, which are much smaller and easier to process and store. Log data is usually text records of flows or events categorized as having a specific significance level. Data collected from security devices for monitoring typically contains a mixture of connection, flow, and state data.

Data storage and prior processing are required to transform the reporting data into a prepared format ready for analysis. In the case of continuous reporting, incoming data usually first arrives in a queue before being transferred to the storage system. Based on the volume and requirements of the reporting data, the data can be stored in a relational database, NoSQL document-based database, or a distributed file system.

Accounting data is generally the easiest to store due to its small size. Depending on the vendor of network condition monitoring equipment, NetFlow accounting data can either be directly stored in an RDBMS or imported into a distributed database. Raw logging data is usually fed into a NoSQL storage cluster for gathering, retention, and processing before storing it in a data warehouse [17].

Equation 2: The confusion matrix and a few performance measures

$$ACC = accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$SP = specificity = \frac{TN}{FP + TN}$$

$$TPR = sensitivity = \frac{TP}{TP + FN}$$

$$FPR = (1 - specificity) = \frac{FP}{FP + TN}$$

3.2. Data Preprocessing and Transformation

With the explosion of data volumes and velocity, organizations are faced with the daunting task of detecting and responding to security threats promptly. Automated and immediate decision-making systems that can adapt to the ongoing evolution of threats are the foremost requirement for the new era of cybersecurity big data applications. However, making real-time decisions is limited by the ability of the detection and verification systems to monitor data of enormous change rate and volume. In addition, models that are currently in place for many big data detection systems are one-off assessments that cannot be updated based on new evidence that is being constantly streamed into the system. A novel cascade architecture of hash functions and miners based on the streaming computing framework is proposed. Hash functions with safety and efficiency features are constructed for the detection of suspected events under big data problems. Data preprocessing and transformation are crucial tasks in the automated detection of threats, such as the detection of abnormal transactions that potentially denote fraud in online services. Data preprocessing consists of selecting the subset of attributes that will be analyzed by the subsequent detection algorithms. Data transformation consists of adjusting the global distribution of the values of the selected attributes so that they conform to an ideal representation or format to be processed by automatic threat detection techniques. The data to be filtered are correlated with several consecutive time intervals, and the filtering is done through a cascade architecture of classifier systems which process the consecutive intervals of filtered data in a parallel fashion. The systems operate as decision nodes and influencers in a more complex detection system which processes the input data in a deluge passing from filters to detectors over probabilities of suspicious events [14].

4. The Role of AI/ML in Cybersecurity Threat Detection

Emerging as a progressive subfield of artificial intelligence, it explores the development of algorithms and models that enable computers to improve their performance in recognizing knowledge and patterns from a new set of data. Employing various techniques and cycles to learn and enhance system performance over time by performing preset tasks in a specific domain. Machine learning techniques are categorized into three primary categories based on the amount of human supervision employed when training the model.

The first category is unsupervised machine learning, described as the process of recognizing patterns from unlabeled data. The second category is supervised machine learning, the process of learning predictive functions using a labeled dataset. Lastly, semi-supervised machine learning combines the two previous techniques to improve machine learning tasks by employing both labeled and unlabeled datasets. The threat detection process in cybersecurity can be divided into four steps: data collection, data preprocessing, model training, and model evaluation. With the increasing volume and complexity of data, it has become impossible to detect cyber threats without automated processes due to the high workload. Machine learning techniques have been widely investigated and utilized in various cybersecurity domains because of their robustness and reliability in threat detection from large volumes of raw data [21].

Deep learning, a representation learning technique based on artificial neural networks, has gained attention as a novel and high-potential technique in various research areas. Characterized by learning distinct data representations with successive layers of abstraction, it automatically determines the appropriate features to improve the performance of the model. The origin of the concept of deep learning can be traced back to the 1980s with the invention of neural networks possessing hidden layers. Despite this early innovation, the field remained stagnant for decades due to the absence of suitable training algorithms, powerful computers, and large datasets. Since 2010, deep learning has resurfaced in research, garnering great interest and achieving impressive performance in various applications. In recent years, substantial investigations have taken place into employing deep learning in a range of cyber threat detection, prediction, and classification tasks due to its efficacy in uncovering concealed patterns in vast datasets.

Continuously feeding data into the model, builds up experience over time about the data and how to detect anomalies and make predictions. With this experience, the model can flag new incoming data that deviate from the expected norm as anomalies. The model can directly classify certain deviations as cyber threat events if they match known threat patterns. Therefore, this AI/ML scheme is a novel and automatic solution for improving efficiency and resource management while maintaining robust and reliable cybersecurity. However, this complex scheme requires additional layers of precaution to avoid misuse. Scientists researching this area need to have a deep understanding of data processing and concerns over accessibility and ethical use; cybercriminals can find and utilize these loopholes to improve their attacks.

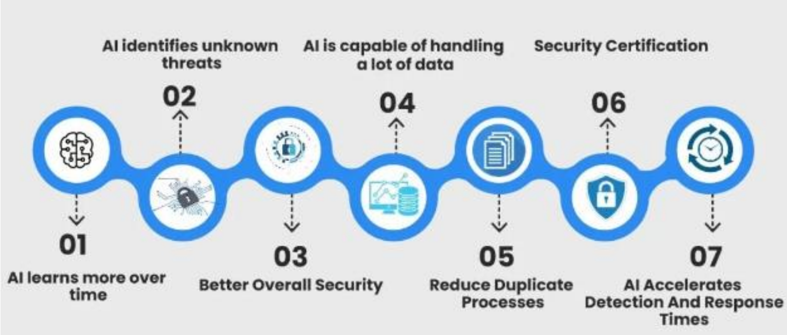

AI identifies unknown threats
AI is capable of handling a lot of data
Security Certification
02
04
06
01
AI learns more over time
03
Better Overall Security
05
Reduce Duplicate Processes
07
AI Accelerates Detection And Response Times

Figure 5. AI and Machine Learning for Threat Detection in Cybersecurity

4.1. Machine Learning Algorithms for Threat Detection

The massive growth of data and the necessity for online services have made organizations more susceptible to cyber threats. One of the most challenging tasks in cybersecurity is the detection of threats through various types of network traffic and system logs. To accomplish this task, machine learning is an inevitable part of both industrial systems and academia. Machine learning methods require constructing computational models based on various data from the treated domains. Each of these methods has its advantages and disadvantages in terms of assumptions made about the data, computational resources used, etc.

Data with the same set of features and the same format is modeled by a specific machine learning algorithm. This learning process is called the training of the model. Once this training phase is accomplished, the work of analyzing the data can proceed. In a given time frame, a set of data is provided to the model; this model then performs a set of calculations on the observed data, the final output being the acceptance or rejection of the considered event.

Machine learning methods can be grouped according to different criteria. One of the most important is the requirement of labeled data – data for which the status of the event is known. If such data is used, the approach is called supervised learning. If not, it is called unsupervised learning. If data is required to be classified into a predefined set of classes, the problem treated is called classification, while if not, the task is called clustering. Another important aspect of machine learning algorithms is the reference made to the time evolution of the model. If the model is always constant, unaltered with the time evolution of the dataset, a static model is described, while if the model is periodically altered as new data is received, a dynamic model is concerned.

Another taxonomy of such methods is done according to the availability of the data. If all the historical data is available since the start of the activity of the system to model, one speaks of an offline approach, while if the data received is processed in real-time, being fed into the model and affecting its output at the same time, an online approach is concerned. In cybersecurity, both supervised and unsupervised methods can be successfully used. The advantage of supervised methods is the possibility of detecting new types of attacks with sufficient precision. The disadvantage is the need for labeled data and the usual fact that labeled data is rare. The advantage of unsupervised methods is the lack of need for labeled datasets. The drawback is the high false alarm rate and the possibility of discriminating against only known attacks.

Machine learning methods can also be classified by their mathematical foundations, for example, support vector machines, artificial neural networks, decision trees, Bayesian networks, and similarity-based methods. Support vector machines originated in computational geometry to separate two classes of data points with a hyperplane by maximizing the margin on both sides of the plane. The main operation required to be performed with this machine learning method is that the database needs to be fitted to the hyperplane. Once the hyperplane is defined, the data can be tested, the classification being accomplished according to the parameters of the output function and its limits.

Another frequently used method in cybersecurity currently is artificial neural networks. A neural network is composed of several layers: the initially presented layer (the input layer), the internal layers – with varying numbers of nodes in different methods, which encode the features of the modeled dataset; and the output layer, which extracts the classification. The nodes of different layers are connected through weights that are calculated in a learning process. Once the model is created, it can classify unknown data or extract the closest class, performing the clustering task.

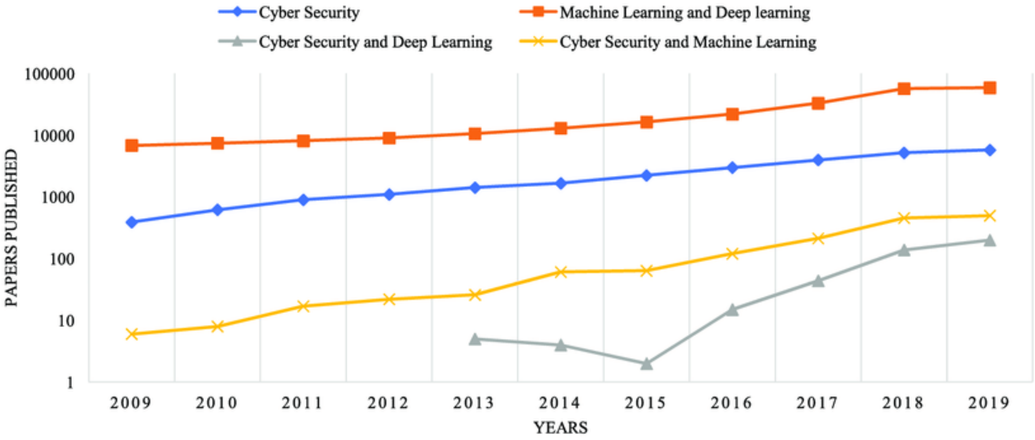


Figure 6. Publications Trends of Machine Learning and Cyber Security

4.2. *Deep Learning Techniques in Cybersecurity*

Deep learning is a branch of artificial intelligence focused on the development and application of learning algorithms inspired by the structure and function of the human brain. Neural networks are the core type of model used in deep learning research. Simple neural networks consist of an input layer of nodes that receives raw data, one or more hidden layers in which the majority of computation occurs, and an output layer of nodes producing the model's predictions or classifying the input data. Each layer is composed of multiple computational nodes. In a supervised learning setup, models are trained using input data and the corresponding ground-truth annotations. The model makes predictions on the input datum, the predictions are compared to the ground truth, and the outgoing activations of all connections to/from the neurons in the various layers are updated according to optimization schemes. Deep learning architectures can address different types of data, i.e., images, text, time series, graphs, etc. This section starts by presenting the most common deep learning models used in cybersecurity and then focuses on graph neural networks, which have gained popularity in recent years.

The most popular neural network architectures in cybersecurity include Convolutional Neural Networks, Recurrent Neural Networks, and graph neural networks. Convolutional Neural Networks are formed by stacking convolutional layers together, where each layer detects different features of the incoming data, transitioning from simple to more complex features. Convolutional Neural Networks are the state of the art for automatically analyzing images and videos and are often the architecture of choice for analyzing raw packet captures in cybersecurity. Recurrent Neural Networks are neural network architectures that have recurrent connections and, thus, a form of memory. Recurrent Neural Networks were originally developed for processing time series or sequential data. In cybersecurity, Recurrent Neural Networks are the architecture of choice when raw network traffic is processed. Graph neural networks are a recently proposed family of models that generalize deep learning principles to irregular graph-structured data. Because of the growing interest in learning from complex data structures, and because of the popularity of attacking and defending real-world network graphs in cybersecurity, graph neural networks represent a promising alternative in this field. Specifically, in terms of learning from complex structures in the real world, phishing detection is an important business problem in cybersecurity, and graph neural networks have been very recently applied to this detection task.

5. Challenges and Future Directions

Although big data and artificial intelligence/machine learning analytics offer promise for enhancing cybersecurity and threat detection, there are several important challenges and future directions. Ethical considerations surrounding data usage present a particular concern in this emerging field. There is an ethical worry that bad actors may exploit the same big data analytics and AI/ML advances for malicious purposes. Furthermore, there is a question concerning the right of individuals and organizations to keep their data, how their data may be analyzed, the types of analysis, and whether they can understand the decisions being rendered. Indeed, it may be impossible for an average person to understand how AI and ML algorithms function and make decisions. Thus, there is concern that it would be difficult for victims of inappropriate data use or breaches to understand how their rights were abused. Similarly, algorithms that reinforce known or unknown biases by inadvertently using biased training data or unjustified features could lead to unfair or damaging outcomes for particular individuals. Organizations collecting data on both citizens and clients are, consequently, subject to ethical scrutiny by various interests. In summary, there is a social responsibility for organizations using big data to be ethical regarding societal concerns raised about abuses or breaches involving the kinds of data being collected, how this data is being processed, the algorithms that are making decisions based on this data, and the outcomes of these decisions.

Another pre-existing and ongoing issue is that of privacy concerns. Big data and AI/ML have the potential to erode personal privacy, introducing a compelling need to find ways to enhance privacy protection in the organizational processing of data. For instance, many organizations collect data on individuals and use the data for various purposes, including the provision of personalized services steeped in privacy-sensitive information. The widespread use of big data by both firms and public institutions may often involve intrusive techniques. Similarly, privacy is a public good because there are factors that allow countries, nation-states, and organizations to socially bilaterally level the playing field regarding the uses and abuses of personal data. Indeed, another issue is ethical that deeply resonates with social concerns about fairness, liability, accountability, disruptions of social and public norms, lack of comprehensive collaboration with the intent to protect public interests, and a lack of transparency.

Additionally, the integration and scalability of threat detection systems integrating big data analytics, AI, and ML with legacy systems pose significant challenges. Further, scalability is a particularly important end-user concern that presents a daunting challenge for data scientists who have little control over the availability and extensiveness of data being gathered. Compounding the challenge of integrating with legacy systems is bundling big data and AI/ML with cybersecurity solutions utilizing compliance-based data management analysis of large amounts of suspicious activity. As a consequence, the systems deployed to comply with regulations often have little in common with the robustness and intelligence enlisted against outside threats by financial firms. Similarly, the design complexity and monitoring costs of new data-driven systems necessitate bundling big data and AI/ML with n-tier systems with demonstrable savings. In addition, current platforms have been externally developed despite the need for compliance assurance with legacy systems.

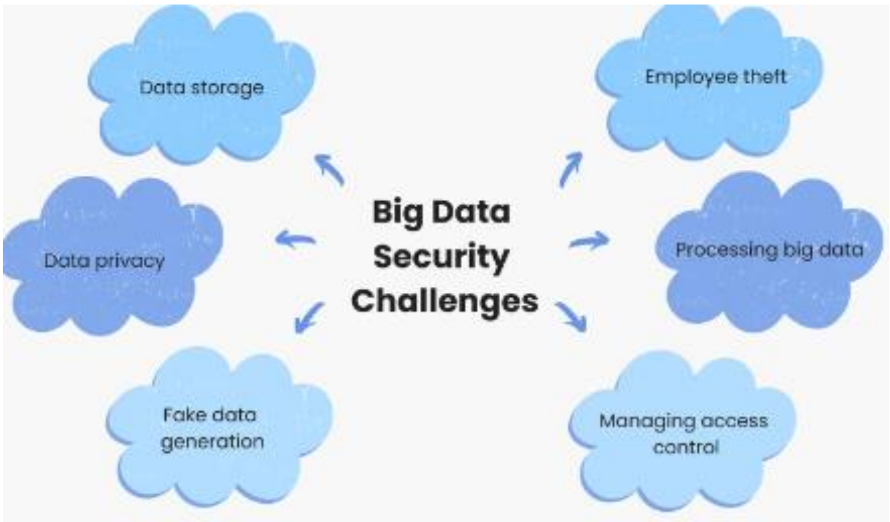


Figure 7. Big Data Security Challenges

5.1. Ethical and Privacy Concerns

Although the potential of Big Data and AI/ML is great, organizations and the public at large must consider ethical and privacy concerns, as both types of technology come with risks. Big Data, due to its nature, raises privacy and security concerns that could negatively impact individual rights if not managed properly. This comes as more and more information is generated about individuals, and that information varies from the very private and intimate to what may be seemingly harmless. Even what may seem like harmless data can have great consequences on individual lives when combined with other such information in the context of Big Data. Organizations undoubtedly benefit from such datasets, as analysis can reveal insights on individual behavior, predict future behavior, and allow firms to operate more effectively as they target specific individuals. These technologies can give organizations unprecedented power over the individuals they govern, as they determine aspects like credit rating, employment eligibility, or propensity to criminality, often without individuals being aware of it or being able to contest it. There is thus a need for more comprehensible regulation concerning the collection and analysis of such datasets and more transparency to boost social trust in organizations benefiting from Big Data analytics.

Concerns of discrimination make regulation particularly difficult. This is aggravated by the inability of affected individuals to understand such discrimination due to the use of complex algorithms that individuals cannot comprehend or contest. If an individual receives a more lucrative credit offer than her neighbors or is subject to more police inquiries than others, is she being discriminated against based on race or ethnicity? Or is this the product of purely economic reasoning that seeks to operate with fewer resources the credit risk that is statistically most likely to default? Complexity in data analysis and modeling, automated decision-making, and the opacity of algorithms are intensifying long standing concerns regarding the fairness of such discrimination.

AI is seen as a double-edged sword that presents ethical challenges. The value judgments surrounding data selection, explanation criteria, and the socio-technical issue of algorithm delegation surface very diverse ethical concerns, such as the de facto exclusion of vulnerable populations from benefits purported by AI systems; the reinforcement of systemic inequalities, and the invisibility of social injustices; the emergence of a history of algorithmic decisions that create past dependency paths; and the lack of accountability among private actors affecting the core of citizens' lives.

Thanks to advances in technology, automated algorithms can access and collect massive amounts of data from any activity, and sensitive personal information is generated and shared, revealing the broadest and contiguous aspects of individuals. Data scientists and knowledge engineers use predictive and causal reasoning to evaluate potential connections between personal information and behavioral/health outcomes, shaping how physicians and citizens access preventative measures, and potentially redesigning the behavior, agency, choices, and goals of individuals. New context-aware objective aggregators create graphs of trust, structural equivalence, and long-term relationships of individuals with others, inferring and classifying their personalities and preferences based on micro-characteristics, as the past is seen as a unique succession of randomly associated events that are not aware of its end. If abuse of such personal information can occur, the behavioral gap of biggest importance is that of the exposed without considering theirs or inquiring negatively influences the rest of the network, potentially revolving feedback loops with irreversible social consequences for individuals [21].

Equation 3. Evaluate the machine learning model for imbalanced data

$$\begin{aligned}
precision &= \frac{TP}{TP+FP}\\
recall &= \frac{TP}{TP+FN}\\
F1 &= \frac{2\times precision \times recall}{precision + recall}\\
accuracy &= \frac{TP+TN}{TP+FN+TN+FP}\\
specificity &= \frac{TN}{TN+FP}
\end{aligned}$$

5.2. Integration and Scalability Challenges

Despite the promise shown by the use of big data and AI/ML techniques to improve threat detection, various integration and scalability challenges exist. Most large companies have a patchwork of different systems to tackle different aspects of security. Building one security system on top of, or instead of, this patchwork would be difficult at best. Designing systems such that there do not have to be intricate integrations or replacements in the patchwork would be much more feasible. Additionally, the scalability of either using only big data for detection or only using AI/ML for detection is highly questionable. Separating attackers from defenders remains a problem regardless of the intricate systems used to scale up systems. Furthermore, there are still various technical challenges to be overcome for the widespread use of big data and AI/ML techniques to improve the monitoring and detection of vulnerabilities, malware, and attackers. There is a large and long-standing gap between detection and response times, especially at the point of incident realization. Improvements at these points heavily rely on automation, which remains a challenge. Additionally, there is a long-standing gap between false positive rates and attack success rates, particularly at the broader and high-level parts of the detection spectrum. Reducing the generation of false positives is critical since they lead to wasting time and resources. Lastly, in the developed detection-amplifying systems, it remains a challenge to reconcile speed and thoroughness, particularly since the assets concerned are located outside the consensus of national boundaries.Despite the potential of big data and AI/ML techniques to enhance threat detection, significant integration and scalability challenges persist, particularly in large organizations that often rely on a patchwork of disparate security systems. Creating a cohesive security framework that overlays or replaces these varied systems proves to be a formidable task. To address this, designing solutions that minimize the need for complex integrations is essential. Moreover, the reliance on either big data or AI/ML in isolation raises questions about scalability and effectiveness. The persistent divide between attackers and defenders remains a critical issue, compounded by technical hurdles that hinder the broader application of these advanced techniques for monitoring and detecting vulnerabilities, malware, and intrusions. The latency in detection and response, especially at the moment of incident realization, underscores the necessity for automation, which continues to face challenges. Furthermore, the ongoing struggle to balance false positive rates with actual attack success rates complicates efforts, as high false positives can drain resources and impede response efforts. Lastly, in developing advanced detection systems, reconciling speed and thoroughness is particularly challenging, especially given that many assets are situated beyond national boundaries, complicating global security efforts [20].

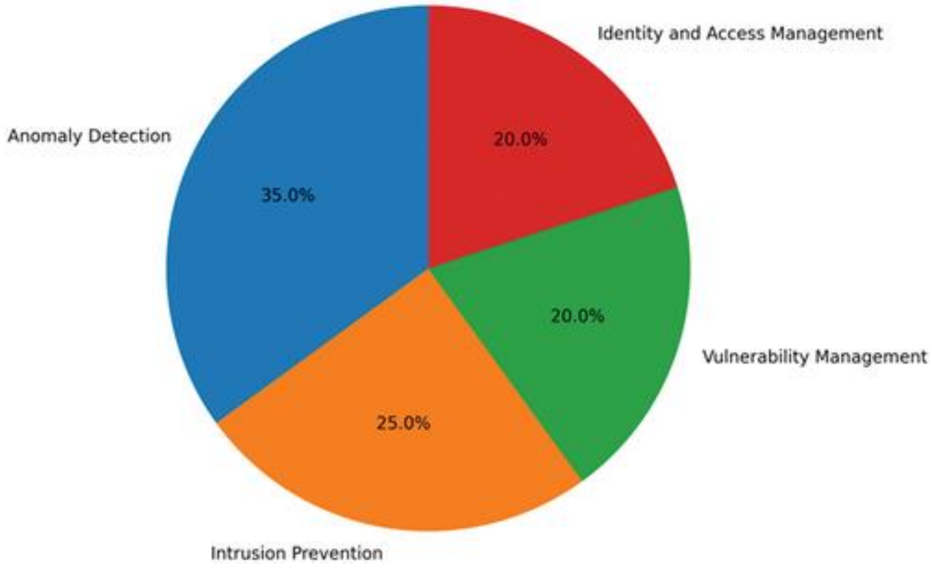


Figure 8. AI/ML Usage in Security Automation

6. Conclusion

In the context of Big Data, cybersecurity encompasses a diverse range of topics being researched. Over the years, the field of computer science has consistently witnessed the development of new tools and methodologies to address emerging challenges. Researchers have already proposed different intelligent techniques for successful log analysis. Experts conduct investigations and build a system to detect intrusion attempts, which conclude that Artificial Neural Network demonstrates the highest accuracy. Moreover, a different sensor is used to detect brute-force attacks on the computer system and find that the k-nearest Neighbor performs better.

With the instant growth of technology and computer systems, vast amounts of data are continuously generated every second. This data is called Big Data. Storing this data becomes cumbersome without proper tools. Moreover, some data are confidential and charge a fine if disclosure happens. A different type of cyber-attack is designed to steal this secret data. Most of these data are stored on remote platforms called cloud storage systems, and many clients share the same storage. Therefore, the investigation of this storage also becomes tedious, and it is necessary to inject intelligent techniques into it for efficient investigation and storage. An engineered and intelligent model is required to detect this multithreaded attack that steals sensitive but unequal data sizes. The proposed model implements the two data mining techniques Data Cleaning and Data Mining on the log packets generated while running different workloads.

The work demonstrates that such important applications can benefit from intelligent techniques and implement Data Cleaning and Data Mining algorithms to detect different cyber-attacks. After applying Data Cleaning, the results show an error percentage of 38% regarding false positives in data mining. This percentage decreases to 1.18% after Data Mining, i.e., the proposed model's accuracy regarding false positives can detect 98.82% of attacks.

6.1. Future Trends

In recent times, the cybersecurity landscape has grown more sophisticated, and advanced technologies are being developed and increasingly adopted in cyber operations. Technologies such as artificial intelligence, machine learning, and big data analysis are redefining the possibilities in the field of cybersecurity for both good and bad. Cybercriminals are now using artificial intelligence and machine learning to perform attacks and cyber operations in a more sophisticated manner and at a larger scale, trying to create more misinformation and operations that are stealthy and evasion-prone. Attackers are also leveraging modern big data technologies to build malicious decision support systems that efficiently gather information, combat security measures, and avoid detection. In the information overload age, the ability to block overwhelming data is at the heart of security management. Hence, security applications must evolve to develop intelligent systems able to transform massive raw data into real knowledge, capable of learning and detecting evolving attacks and threats. Therefore, in the future, it is expected that offensive cyber operations will use machine learning as a core technology, and defense technologies must thoroughly explore and adopt the use of machine learning to prevent nations.

Artificial intelligence and machine learning technologies are likely to play a vital role in the future of cyber defense. Automated threat detection and investigation, motivated by safety incidents or other business needs, will be based on the automation of security tasks that rely entirely on the behavior and capabilities of artificial intelligence and machine learning technologies, rather than being assisted by humans. To address the challenge of detecting unknown attacks, the ability to learn and classify complex data with sophisticated machine-learning techniques will increasingly be part of cyber defense systems. As awareness grows of the importance of dealing with complex attacks that abuse the security structures of existing monitoring systems, there will also be a growing interest in the understanding and modeling of these monitoring systems, their security assumptions, and their vulnerabilities.

The rapid adoption of software-as-a-service solutions, cloud infrastructures, and on-cloud storage and processing holds great opportunities but also poses further cybersecurity challenges. Increasingly, analysis must be performed outside the traditional enterprise, which leads to the growing interest in analysis-as-a-service solutions. Similar to modern antivirus detection engines, this will lead to a greater need for simple client-side tools capable of efficiently communicating with remote detection engines. Typically, this entails a growing interest in off-the-shelf monitored protocols and technologies to conduct such communication without being hampered by security architectures or adverse network behavior. Additionally, as multiple applications often share similar datasets, the risk of data poisoning attacks targeting datasets must be considered, as well as data privacy issues.

References

[1] Pamulaparthyvenkata, S. (2022). Unlocking the Adherence Imperative: A Unified Data Engineering Framework Leveraging Patient-Centric Ontologies for Personalized Healthcare Delivery and Enhanced Provider-Patient Loyalty. Distributed Learning and Broad Applications in Scientific Research, 8, 46-73.

[2] Mahida, A. (2022). Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-249. DOI: doi. org/10.47363/JAICC/2022 (1), 232, 2-4.

[3] Avacharmal, R. (2022). ADVANCES IN UNSUPERVISED LEARNING TECHNIQUES FOR ANOMALY DETECTION AND FRAUD IDENTIFICATION IN FINANCIAL TRANSACTIONS. NeuroQuantology, 20(5), 5570.

[4] Pamulaparthyvenkata, S., & Avacharmal, R. (2021). Leveraging Machine Learning for Proactive Financial Risk Mitigation and Revenue Stream Optimization in the Transition Towards Value-Based Care Delivery Models. African Journal of Artificial Intelligence and Sustainable Development, 1(2), 86-126.

[5] Yadav, P. S. (2021). Improving DevOps Efficiency with Jenkins Shared Libraries and Templates. European Journal of Advances in Engineering and Technology, 8(11), 116-120.

[6] Perumal, A. P., Deshmukh, H., Chintale, P., Desaboyina, G., & Najana, M. Implementing zero trust architecture in financial services cloud environments in Microsoft azure security framework.

[7] Vaka, D. K. "Artificial intelligence enabled Demand Sensing: Enhancing Supply Chain Responsiveness.

[8] Tilala, M., Pamulaparthyvenkata, S., Chawda, A. D., & Benke, A. P. Explore the Technologies and Architectures Enabling Real-Time Data Processing within Healthcare Data Lakes, and How They Facilitate Immediate Clinical Decision-Making and Patient Care Interventions. European Chemical Bulletin, 11, 4537-4542.

[9] Mahida, A. (2022). A Comprehensive Review on Ethical Considerations in Cloud Computing-Privacy Data Sovereignty, and Compliance. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-248. DOI: doi. org/10.47363/JAICC/2022 (1), 231, 2-4.

[10] Manukonda, K. R. R. Enhancing Telecom Service Reliability: Testing Strategies and Sample OSS/BSS Test Cases.

[11] Avacharmal, R., & Pamulaparthyvenkata, S. (2022). Enhancing Algorithmic Efficacy: A Comprehensive Exploration of Machine Learning Model Lifecycle Management from Inception to Operationalization. Distributed Learning and Broad Applications in Scientific Research, 8, 29-45.

[12] Yadav, P. S. (2021). Big Data Analytics and Machine Learning: Transforming Fixed Income Investment Strategies. North American Journal of Engineering Research, 2(2).

[13] Perumal, A. P., & Chintale, P. Improving operational efficiency and productivity through the fusion of DevOps and SRE practices in multi-cloud operations.

[14] Vaka, D. K. " Integrated Excellence: PM-EWM Integration Solution for S/4HANA 2020/2021.

[15] Mahida, A. A Review on Continuous Integration and Continuous Deployment (CI/CD) for Machine Learning.

[16] Manukonda, K. R. R. (2022). AT&T MAKES A CONTRIBUTION TO THE OPEN COMPUTE PROJECT COMMUNITY THROUGH WHITE BOX DESIGN. Journal of Technological Innovations, 3(1).

[17] Avacharmal, R. (2021). Leveraging Supervised Machine Learning Algorithms for Enhanced Anomaly Detection in Anti-Money Laundering (AML) Transaction Monitoring Systems: A Comparative Analysis of Performance and Explainability. African Journal of Artificial Intelligence and Sustainable Development, 1(2), 68-85.

[18] Manukonda, K. R. R. Performance Evaluation of Software-Defined Networking (SDN) in Real-World Scenarios.

[19] Chintale, P. (2020). Designing a secure self-onboarding system for internet customers using Google cloud SaaS framework. IJAR, 6(5), 482-487.

[20] Vaka, D. K. (2020). Navigating Uncertainty: The Power of 'Just in Time SAP for Supply Chain Dynamics. Journal of Technological Innovations, 1(2).

[21] Mahida, A. A Comprehensive Review on Generative Models for Anomaly Detection in Financial Data.

[22] Dilip Kumar Vaka. (2019). Cloud-Driven Excellence: A Comprehensive Evaluation of SAP S/4HANA ERP. Journal of Scientific and Engineering Research. https://doi.org/10.5281/ZENODO.11219959

[23] Chintale, P. SCALABLE AND COST-EFFECTIVE SELF-ONBOARDING SOLUTIONS FOR HOME INTERNET USERS UTILIZING GOOGLE CLOUD'S SAAS FRAMEWORK