International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1440-1448
https://doi.org/10.70917/ijcisim-2026-5703
Article

BIG DATA ANALYTICS FOR CYBERSECURITY THREAT DETECTION: A REVIEW

Kushwaha Divya Dharampal Singh[1], Sonakshi Priya[2], Saurabh Kumar[3], Danish Shamshee[4], Navnit Kumar Maurya[5], Sonali Sagar Kharade[6]

[1,2,3,4,5]Parul Institute of Engineering and Technology, MCA Department, Vadodara, Gujarat.
[1]Email: divyakushwaha0807@gmail.com
[2] Email: sonakshipriya2031@gmail.com
[3] Email: brsaurabhkr@gmail.com
[4] Email: danishshamshee@gmail.com
[5] Email: navneetmaurya716@gmail.com
[6] Assistant Professor, Parul Institute of Engineering and Technology, MCA Department, Vadodara, Gujarat.
Email: kharadesonali1234@gmail.com
Corresponding Author: Kushwaha Divya Dharampal Singh

Abstract: Modern digital ecosystems built on cloud platforms, the Internet of Things (IoT), and artificial intelligence have made cyberattacks both more frequent and more complex. Conventional security tools increasingly struggle to detect and respond to sophisticated intrusions as they unfold. Big Data Analytics (BDA) has emerged as a leading response to this gap, offering the capacity to process enormous volumes of structured and unstructured security data in order to surface hidden patterns, anomalies, and emerging threats. This review examines how the cybersecurity threat landscape has changed and how big data analytics contributes to stronger detection and response capabilities. It considers the defining traits of big data — volume, velocity, variety, veracity, and value — and surveys the technology stack commonly used to manage it, including Hadoop, Apache Spark, Kafka, Flink, Hive, HBase, and various NoSQL systems. The paper further examines how machine learning, when paired with big data infrastructure, enables real-time intrusion detection, anomaly identification, and predictive security analytics. Persistent obstacles are also discussed, among them advanced persistent threats, IoT-specific weaknesses, AI-enabled attacks, and cloud security gaps. The review closes by summarizing recent technical progress and arguing for scalable, intelligent, data-driven security architectures capable of safeguarding today's digital infrastructure against an evolving threat environment.
Keywords: Cybersecurity, Big Data Analytics, Internet of Things, Artificial Intelligence, Cyber Threats

1. Introduction

Digitization has been an ongoing and continually expanding process, largely valued for removing manual effort and boosting operational efficiency. Jang-Jaccard and Nepal observe that society, the economy, and critical infrastructure have all grown heavily reliant on technology **[13]**. Siemens' 2018 cybersecurity report similarly noted that as billions of devices become interconnected through the Internet of Things, the digital world is being reshaped at every level — a shift that carries enormous opportunity alongside substantial risk **[11]**. As more of our critical infrastructure, economic activity, and social life migrates into cyberspace, this same connectivity creates fertile ground for malicious actors to exploit. Jang-Jaccard and Nepal further note that growing dependence on information technology gives attackers greater incentive and the potential to cause severe damage **[13]**. Key areas of concern

include critical infrastructure, cybercrime, the cultural dimensions of technology use, and the broader set of threats and challenges these trends produce.

Organizations worldwide now treat cybersecurity as a pressing priority, a shift driven by threats that have grown steadily more complex and frequent over recent decades. Early cyberattacks were comparatively simple — viruses and worms that spread with limited sophistication, typically targeting individual machines or small networks to disrupt operations or steal data **[32]**. As digital systems have become woven into the fabric of business operations, however, attacks have kept pace, evolving into multifaceted campaigns capable of threatening organizations of any size or sector **[14]**. This convergence of technology and everyday life has also turned cybercrime into a persistent economic burden, with attacks increasingly threatening organizational continuity, financial stability, and public trust **[7]**. Industry estimates further indicate that the total volume of data generated worldwide continues to expand at a rapid pace, a trend illustrated in Figure 1, which in turn magnifies both the opportunity and the difficulty of applying analytics to cybersecurity at scale.

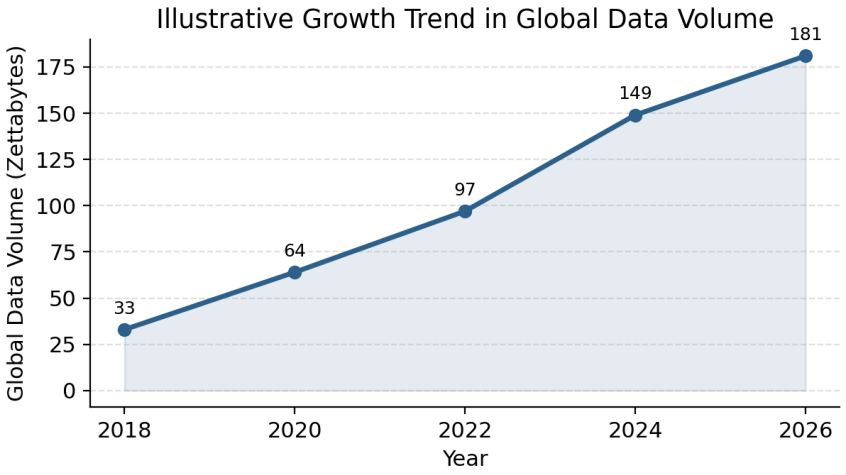

Figure 1. Illustrative Growth Trend in Global Data Volume (2018-2026, industry estimates)

2. Types of Cybersecurity

- Critical infrastructure security: Legacy software in SCADA (supervisory control and data acquisition) systems often leaves organizations that run essential services more exposed to attack than typical businesses. In the UK, the NIS Regulations require operators of energy, transport, healthcare, water, and digital infrastructure services — as well as digital service providers — to adopt adequate administrative and technical safeguards against information security risks [17].
- Network security: This covers vulnerabilities across operating systems and network architecture — servers, hosts, firewalls, wireless access points, and communication protocols — and the measures used to close them.
- Cloud security: Protects data, applications, and infrastructure hosted in cloud environments.
- IoT security: Concerns the protection of internet-connected smart devices — including fire alarms, lighting systems, thermostats, and other appliances — that operate without direct human input.
- Application security: Addresses vulnerabilities introduced through insecure development practices during the design, coding, and deployment of software or websites; also referred to as application protection [7].

Table 1 summarizes these categories alongside their primary focus area and a representative example, for quick reference.

Category	Primary Focus	Representative Example
Critical infrastructure security	Protecting essential-service operators (energy, water, health)	SCADA system hardening
Network security	Securing network architecture and traffic	Firewalls, intrusion detection
Cloud security	Protecting cloud-hosted data and services	Access control, encryption
IoT security	Securing connected smart devices	Device authentication
Application security	Removing flaws from software/web development	Secure code review

3. Current Challenges in Threat Detection

Cybersecurity remains a pressing concern for individuals, businesses, and governments in the digital era. As reliance on technology and connected devices grows, protecting systems, networks, and data from unauthorized access, theft, and damage has become increasingly urgent. Despite technological progress, safeguarding organizations, their employees, and their critical assets continues to present significant obstacles.

3.1 Sophisticated Nature of Cyberattacks

The growing complexity of attacks aimed at gaining unauthorized access to computer systems and networks represents one of the field's central challenges **[29]**. Attackers continuously refine sophisticated techniques to breach defenses or exploit weaknesses, with multi-vector attacks, polymorphic and fileless malware, zero-day exploits, and advanced persistent threats among the most difficult to counter **[26]**.

- Advanced Persistent Threats (APTs): These are highly organized, targeted campaigns aimed at diplomatic bodies, IT firms, military services, chemical industries, and other sensitive sectors, designed to exfiltrate confidential data and inflict lasting damage through sustained, persistent access [8]. APTs are frequently state-sponsored or backed by organized criminal groups seeking economic, political, or strategic gains by targeting national defense, government institutions, manufacturing, military planning, and other high-value areas [10].

3.2 Internet of Things (IoT) Security

The IoT is a rapidly expanding technology that links billions of devices to the internet, connecting sensors and computing hardware through shared protocols to exchange and share data **[9]**. This scale of interconnection introduces serious cybersecurity concerns rooted in the inherent vulnerabilities of IoT systems **[5]**.

IoT devices routinely collect and transmit large volumes of sensitive, often personal, data, making privacy protection essential. Weak encryption, insecure storage, and poor data management practices can open the door to breaches and unauthorized access **[26]**. Because these devices typically rely on wireless protocols such as Wi-Fi, Bluetooth, or cellular networks, attackers can target the communication channel itself through eavesdropping, interception, or tampering. Poorly configured networks and inadequate encryption further compound these risks **[2]**.

3.3 AI-Driven Attacks

Rapid technological advancement and expanding digitalization across sectors have made cybersecurity progressively more difficult to manage. Artificial intelligence and machine learning represent one such advancement, and their misuse has become an emerging attack vector. While AI offers clear security benefits, it also carries a negative impact in the form of AI-driven attacks. These generally fall into two categories: AI-assisted attacks, where machine learning tools help human attackers plan or carry out intrusions, and autonomous AI attacks, where AI agents execute attacks against target systems with no human involvement at all.

3.4 Cloud Computing

Growing reliance on cloud security and cloud services introduces new categories of risk for organizations. Common challenges include data breaches, unauthorized access, insecure APIs, and shared infrastructure vulnerabilities, all of which can lead to data loss and service disruption that leaves critical services unavailable, alongside a range of other security threats organizations must now manage **[30]**.

4. Big Data Analytics

The term "Big Data" describes datasets that have grown too large and complex for conventional database systems to handle efficiently — that is, data whose scale exceeds what standard software and storage tools can capture, manage, and process within a reasonable timeframe **[15]**.

The scale of big data continues to grow, with individual datasets now ranging from tens of terabytes to multiple petabytes. This growth brings persistent challenges around capture, storage, search, sharing, analysis, and visualization. Organizations increasingly mine large, detailed datasets to uncover insights that were previously invisible to them **[23]**.

Big data analytics refers to the application of advanced analytical methods to these large datasets. Analysis performed on large-scale data can reveal meaningful business insight and drive change, though managing that data becomes progressively harder as its volume increases **[23]**.

5. Characteristics of Big Data

Big data is commonly described through the "5Vs" framework **[3]**: volume, velocity, variety, veracity, and value **[1]**, **[19]**.

- Volume refers to the sheer quantity of data involved, typically the first characteristic that comes to mind. Many organizations accumulate vast archives of log data without the means to manage it effectively; the ability to convert this scale into usable information is central to the value of big data analytics [1]. While large volumes offer real business benefit, they also slow down retrieval and analysis due to the computational demands involved [12].

- Velocity describes both the speed at which data is generated and the pace required to manage, store, and analyze it. Hundreds of hours of video are uploaded to YouTube and over 200 million emails are sent through Gmail every single second.

- Variety captures the diversity of data types and sources. Recognizing the category to which a given dataset belongs is essential, since big data is rarely structured and does not fit neatly into relational databases. With roughly 90% of generated data being unstructured, managing this mix of structured and unstructured formats adds considerably to the analytical burden.

- Veracity concerns the reliability of the data itself, which directly affects analytical precision. Given the volume, velocity, and variety involved, guaranteeing complete accuracy across all data is unrealistic; data quality varies, and analytical precision depends heavily on the trustworthiness of the source.

- Value reflects how meaningful or useful the data actually is — arguably the most consequential of the five characteristics. Big data holds substantial potential value, but that value can only be realized through appropriate access and analysis [3].

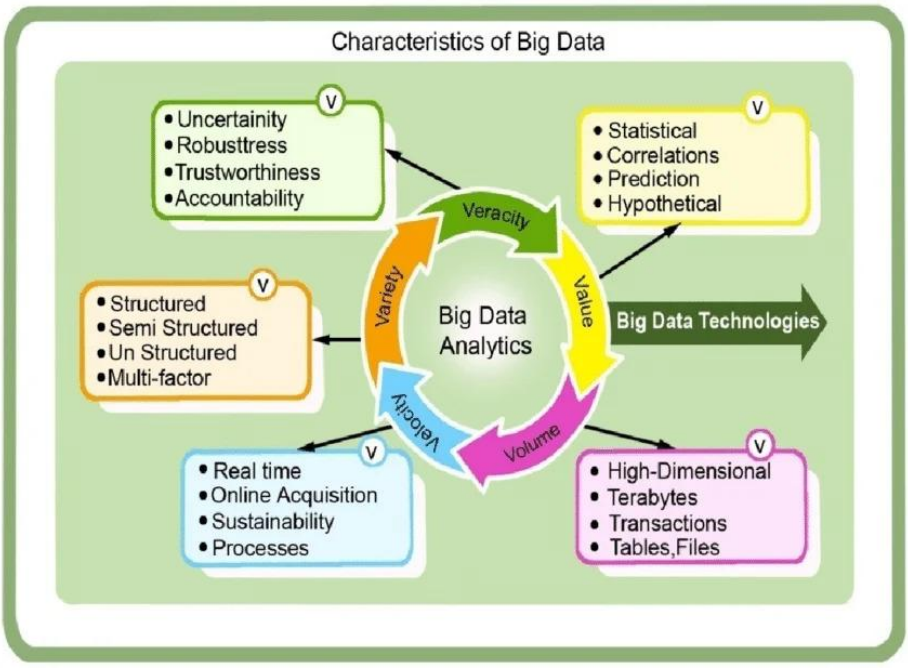

Figure 2. Characteristics of Big Data Analytics [31]

6. Role of Big Data Analytics in Cybersecurity

Organizations adopt big data analytics for many reasons, not least the fresh approaches it brings to broader business analytics. Conventional security technologies are often unable to fully identify threats, fraud, or attacks on their own. A wide range of big data analytics frameworks and tools now exist to detect anomalous activity within massive datasets, with machine learning techniques — including clustering, data visualization, and outlier detection — making it considerably easier to uncover meaningful connections within that data **[16]**, **[27]**.

Big data analytics has been studied and deployed across numerous industry sectors. Its value in cybersecurity specifically has grown alongside increasingly complex networks and increasingly sophisticated threats. Because large volumes of data can obscure the true nature of a threat, cybersecurity stands out as one of the domains where this challenge matters most.

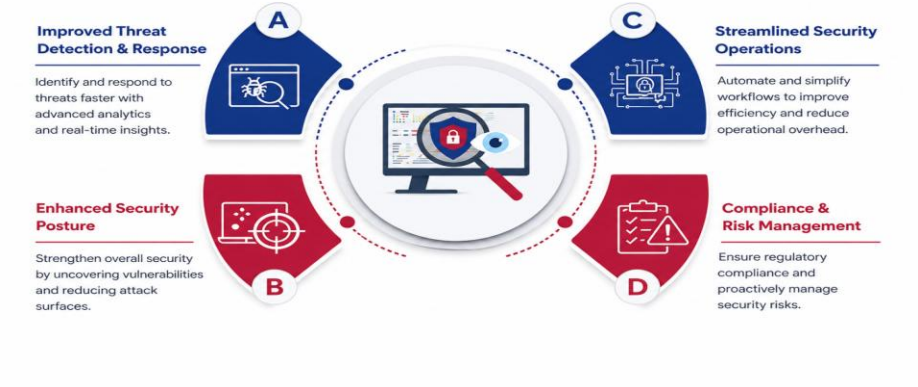

Figure 3. Data Analytics in Cybersecurity [18]

7. Recent Big Data Analytics Tools and Technologies

Tools and techniques for big data analytics have advanced considerably to keep up with the data volumes generated by modern technology. Some of the most widely used are outlined below **[28]**, **[24]**, **[21]**:

- Hadoop: An open-source Apache platform for storing and processing large-scale data across distributed computing clusters, built around the Hadoop Distributed File System (HDFS) for storage and the MapReduce model for processing.
- Spark: A fast, general-purpose cluster computing engine offering in-memory processing, used for batch processing, graph analytics, machine learning, and real-time stream processing.
- NoSQL Databases: Systems such as Cassandra, Couchbase, and MongoDB handle large volumes of semi-structured or unstructured data efficiently, with the flexibility and scalability that big data applications require.
- Apache Kafka: A distributed streaming platform commonly used to build real-time data pipelines, enabling analysis of large data streams as they are ingested.
- Apache Flink: A stream-processing framework built for low-latency, high-throughput processing of continuous data, supporting stateful computation, exactly-once semantics, and event-time processing.
- Apache Hive: A Hadoop-based data warehousing tool that lets users query and analyze large HDFS-stored datasets through a familiar SQL-like interface.
- Apache HBase: A distributed, scalable NoSQL database built on Hadoop that supports random read/write access and can manage tables containing billions of rows.
- Apache Storm: A distributed, fault-tolerant system for real-time stream processing with minimal latency.
- Apache Drill: A distributed SQL query engine that enables interactive analysis across data sources including Hadoop, NoSQL databases, and cloud storage.
- Machine Learning Libraries: Tools such as TensorFlow, scikit-learn, and PyTorch support the development and deployment of machine learning models at scale, integrating with Spark and other big data systems for distributed training and inference.
- Data Visualization Tools: Platforms such as Tableau, Power BI, and Apache Superset allow users to build interactive dashboards and share insights drawn from large datasets.

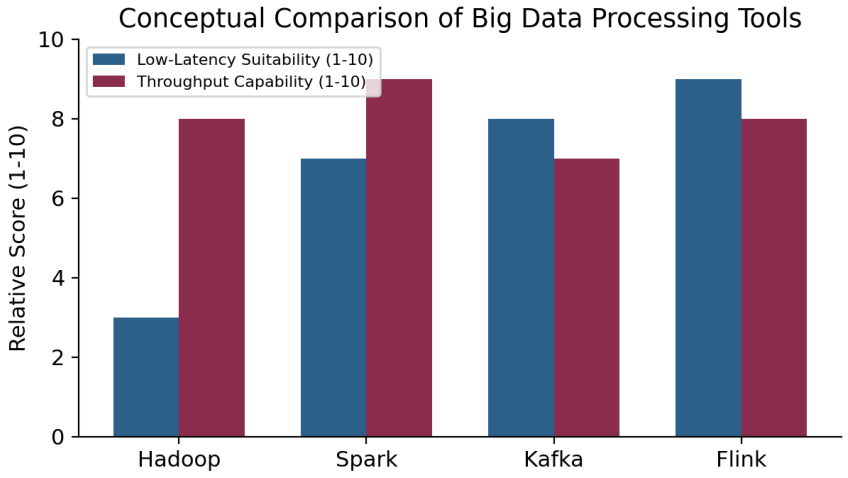

Figure 4. Conceptual Comparison of Big Data Processing Tools (illustrative, not benchmark data)

Table 2 summarizes each tool's primary role and typical application within a cybersecurity analytics pipeline.

Tool	Primary Function	Typical Cybersecurity Use
Hadoop (HDFS + MapReduce)	Distributed storage and batch processing	Archiving and batch analysis of log data
Apache Spark	In-memory cluster computing	Fast batch and ML-based threat analysis
Apache Kafka	Distributed event streaming	Real-time ingestion of security event data
Apache Flink	Low-latency stream processing	Real-time anomaly and intrusion detection
Apache Hive	SQL-like data warehousing on HDFS	Querying large historical security datasets
Apache HBase	Distributed NoSQL database	Fast lookups across billions of log records
NoSQL (Cassandra, MongoDB)	Flexible semi/unstructured storage	Storing varied threat-intelligence data

8. Integration of Machine Learning and Big Data Analytics

Combining machine learning with big data analytics represents a substantial shift in how real-time cybersecurity threats are detected, strengthening security systems by drawing on the complementary strengths of both technologies **[6]**, **[20]**. Machine learning's capacity to learn from and adapt to new data, combined with big data's ability to process massive information volumes, forms a robust framework for countering evolving cyber threats.

Large datasets are essential to effective machine learning: the core value of ML in cybersecurity lies in its ability to analyze extensive data to detect patterns, anomalies, and potential threats **[4]**, **[25]**. Model accuracy and robustness generally improve as more data becomes available. Supervised learning approaches, for example, rely on labeled training data and can achieve high accuracy in threat detection when given extensive, varied datasets **[33]**, learning to recognize patterns tied to known threats and generalizing that knowledge to new instances. Unsupervised methods, which detect patterns without predefined labels, similarly benefit from large datasets by revealing previously unknown anomalies and emerging threat patterns **[22]**. Big data analytics supports the continual refinement of these models with fresh, diverse data, improving both their accuracy and predictive power over time.

9. Future Research Directions

The future of cybersecurity will depend heavily on developing intelligent, adaptive, and scalable big data analytics frameworks capable of processing ever-growing volumes of security data in real time. Future work should prioritize explainable AI (XAI) models that make machine learning-based threat detection more transparent and interpretable. Federated learning and privacy-preserving analytics also hold significant promise for enabling collaborative cybersecurity efforts without exposing sensitive organizational data. Lightweight analytics models tailored to resource-constrained IoT and edge environments are similarly needed to support efficient real-time detection at scale. Emerging technologies such as blockchain, digital twins, and quantum computing warrant further investigation for their potential to strengthen cyber resilience and secure distributed systems. Additional research should focus on building standardized cybersecurity datasets, improving cross-platform interoperability, lowering false positive rates, and strengthening automated incident response. Together, big data analytics, artificial intelligence, and cloud-native security solutions are likely to play a defining role in the next generation of cybersecurity systems built to counter increasingly sophisticated attacks.

10. Conclusion

The accelerating digitalization of businesses, governments, and critical infrastructure has made cybersecurity one of today's most pressing global concerns. As cyber threats grow more sophisticated and far-reaching, traditional security approaches are increasingly unable to keep pace with timely detection and mitigation. Big Data Analytics has emerged as a powerful means of processing vast, heterogeneous volumes of security data and surfacing malicious activity through advanced analytical methods. Pairing machine learning with big data technologies further strengthens the ability to detect anomalies, anticipate attacks, and automate threat response with greater accuracy. Platforms such as Hadoop, Apache Spark, Kafka, Flink, and various NoSQL databases now provide the scalable infrastructure needed to manage large-scale cybersecurity data. Even so, challenges around data quality, privacy, scalability, computational complexity, and constantly evolving attack methods persist. Sustained research and innovation across artificial intelligence, big data analytics, and intelligent cybersecurity frameworks will remain essential to building resilient, proactive, and adaptive systems capable of protecting digital ecosystems against future threats.

REFERENCES

1. Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and Hadoop technology. Procedia Computer Science, 48, 319-324.
2. AsSadhan, B., & Moura, J. M. (2014). An efficient method to detect periodic behavior in botnet traffic by analyzing control plane traffic. Journal of Advanced Research, 5(4), 435-448.
3. Basim Alwan, H., & Ku-Mahamud, K. R. (2020, February). Big data: Definition, characteristics, life cycle, applications, and challenges. In IOP Conference Series: Materials Science and Engineering (Vol. 769, No. 1, p. 012007). IOP Publishing.
4. Bertoni, G., Gritti, S., & Rechberger, C. (2022). Post-quantum cryptography: The new era of secure communication. IEEE Transactions on Information Forensics and Security, 17(2), 45-58.
5. Clark, J., & Van Oorschot, P. C. (2013, May). SoK: SSL and HTTPS: Revisiting past challenges and evaluating certificate trust model enhancements. In 2013 IEEE Symposium on Security and Privacy (pp. 511-525). IEEE.
6. Coventry, L., & Branley, D. (2018). Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas, 113, 48-52.
7. Fadziso, T., Thaduri, U. R., Dekkati, S., Ballamudi, V. K. R., & Desamsetti, H. (2023). Evolution of the cyber security threat: an overview of the scale of cyber threat. Digitalization & Sustainability Review, 3(1), 1-12.
8. Ghafir, I., & Prenosil, V. (2014). Advanced persistent threat attack detection: an overview. International Journal of Advanced Computer Network and Security, 4(4), 50-54.
9. Gunduz, M. Z., & Das, R. (2020). Cyber-security on smart grid: Threats and potential solutions. Computer Networks, 169, 107094.
10. Hejase, H. J., Fayyad-Kazan, H. F., & Moukadem, I. (2020). Advanced persistent threats (APT): an awareness review. Journal of Economics and Economic Education Research, 21(6), 1-8.
11. Hussain, A., Mohamed, A., & Razali, S. (2020, March). A review on cybersecurity: Challenges & emerging threats. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security (pp. 1-7).
12. Ianni, M., Masciari, E., Mazzeo, G. M., Mezzanzanica, M., & Zaniolo, C. (2020). Fast and effective big data exploration by clustering. Future Generation Computer Systems, 102, 84-94.
13. Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. Journal of Computer and System Sciences, 80(5), 973-993.
14. Khan, S., Alhazmi, Y., Khan, F., Ullah, F., & Alanazi, E. (2022). Cybersecurity issues and challenges for IoT-based smart grid applications: A comprehensive analysis. Energy Reports, 8, 9828-9841.
15. Kubick, W. R. (2012). Big data, information and meaning. Applied Clinical Trials, 21(2), 26.
16. Mahmood, T., & Afzal, U. (2013, December). Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In 2013 2nd National Conference on Information Assurance (NCIA) (pp. 129-134). IEEE.
17. Mandapuram, M., Gutlapalli, S. S., Reddy, M., & Bodepudi, A. (2020). Application of artificial intelligence (AI) technologies to accelerate market segmentation. Global Disclosure of Economics and Business, 9(2), 141-150.
18. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., ... & McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.
19. Monino, J. L., & Sedkaoui, S. (2016). Big data, open data and data development (Vol. 3). John Wiley & Sons.
20. Raghavan, S., Kalra, S., & Zhang, Y. (2022). Compliance and privacy challenges in big data analytics: A review. Journal of Privacy and Confidentiality, 13(2), 1-21.
21. Rassam, M. A., Maarof, M., & Zainal, A. (2017). Big data analytics adoption for cybersecurity: A review of current solutions, requirements, challenges and trends. Journal of Information Assurance & Security, 12(4).
22. Ribeiro, S., & Santos, J. (2020). Unsupervised machine learning for anomaly detection in cybersecurity: A survey. IEEE Transactions on Network and Service Management, 17(1), 458-471.
23. Russom, P. (2011). Big data analytics. TDWI Best Practices Report, Fourth Quarter, 19(4), 1-34.
24. Savas, O., & Deng, J. (Eds.). (2017). Big data analytics in cybersecurity. CRC Press.

25. Sengupta, S., Kaushik, S., Krishnamurthy, P., & Kambhampati, S. (2020). A survey of deep learning techniques for cybersecurity intrusion detection. IEEE Communications Surveys & Tutorials, 22(3), 1646-1670.
26. Sharma, R. (2012). Study of latest emerging trends on cyber security and its challenges to society. International Journal of Scientific & Engineering Research, 3(6), 1.
27. Singh, A. (2021). Big data technologies for cyber security in the digital era. Available at SSRN 4629044.
28. Srivastava, N., & Jaiswal, U. C. (2019, March). Big data analytics technique in cyber security: a review. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 579-585). IEEE.
29. Tankard, C. (2011). Advanced persistent threats and how to monitor and deter them. Network Security, 2011(8), 16-19.
30. Thakur, K., Qiu, M., Gai, K., & Ali, M. L. (2015, November). An investigation on cyber security threats and security models. In 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing (pp. 307-311). IEEE.
31. VVR, D. M. R., & Silpa, N. (2015). A comprehensive study on potential research opportunities of big data analytics to leverage the transformation in various key domains. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 5(5), 01-18.
32. Wang, Z., Ma, J., Wang, X., Hu, J., Qin, Z., & Ren, K. (2022). Threats to training: A survey of poisoning attacks and defenses on machine learning systems. ACM Computing Surveys, 55(7), 1-36.
33. Xia, Y., Yang, X., & Zhang, H. (2021). A survey on machine learning for cybersecurity: Current status and future directions. IEEE Transactions on Emerging Topics in Computing, 9(1), 56-68.