Artistic rendition of "A Literature Review of Big Data for Cybercrime Detection"

Dharmraj Kumar Vitragi[1st], Dr. Lalan Kumar Singh[2nd], Arif Mohammad Sattar[3rd], Mritunjay Kr. Ranjan[4th*]

1st Research Scholar, P.G. Dept of Mathematics and Computer Science, Magadh University,
Bodh-Gaya, Bihar, India,
2nd Associate Professor, Dept. of Mathematics, K.S.M. College, Aurangabad, Bihar - India,
3rd Assistant Professor, Dept of Computer Science, A.M. College, Gaya, Bihar - India,
*4th, Assistant Professor, SOCSE - Sandip University, Nashik, Maharashtra- India

Email: [1st]vitragiparaiya@gmail.com, [2nd] drlalankumarsingh1958@gmail.com, [3rd] amsattargaya@gmail.com,
[*4th] mritunjaykranjan@gmail.com

Abstract: Cybercrime has grown, threatening online users' security and privacy. Researchers and practitioners use big data analytics to detect and combat complex cyber threats. This survey of big data analytics research on cybercrime detection discusses recent advances, strategies, and challenges. Cybercrime and big data analytics are introduced. Detecting cybercrime involves spotting irregularities, anticipating attacks, and identifying malicious actions. Investigators examine how cybercrime investigators collect massive amounts of data. Experts improved detection using structured (server logs, network traffic, user activity logs) and unstructured (social media posts, dark web forums) data. Big data analytics-based cybercrime detection is discussed next in the literature review. Supervised, unsupervised, and semi-supervised machine learning are popular because they can handle large amounts of data and detect cybersecurity threats. Big data analytics for cybercrime detection are also studied. It examines how these networks handle viral spread and identity theft. Big data analytics has pros and cons, according to this literature review. Academics and businesses face challenges in data privacy and security, analytics platform scalability, and cyber threat evolution. Results and future research on big data analytics for cybercrime detection are discussed. Academics, industry, and governments must collaborate to develop more effective, scalable, and privacy-preserving cybercrime solutions. Big data analytics has transformed cybercrime detection. It may guide researchers, practitioners, and policymakers.

***Keywords*:** Cybercrime, Big-data, Data-analytics, Cyberattack, Data Privacy, Cyber Detection, Threats

1. INTRODUCTION

The prevalence of digital technology and the ease of global communication have contributed to a meteoric rise in cybercrime. Data breaches, identity theft, and complex cyber-attacks on essential infrastructure all pose serious risks to people, businesses, and governments worldwide. The methods used by cybercriminals are constantly evolving alongside the digital landscape, making it more important than ever to develop novel and efficient strategies for detecting and mitigating them. In this review, incorporating Big Data analytics has shown promise for bolstering cyber defences [11]. The wealth of information contained in Big Data, which is characterised by the exponential growth of data volumes, velocity, variety, and veracity, can be used to improve cybercrime detection and response. To better protect our digital assets and personal information, a new and exciting field has emerged at the intersection of Big Data and cybersecurity [12]. This study will delve deeply into the relationship between Big Data analytics and the identification of cybercrime. It explores the varied terrain of theory and application, illuminating the most salient trends, methodologies, obstacles, and opportunities in this field. Our goal in conducting this literature review is to help researchers, practitioners, and policymakers better understand the current state of affairs and the potential for future advancements by providing a nuanced understanding of the body of knowledge that has accumulated in this field. This review seeks to answer key questions about the use of Big Data to improve cybercrime detection by conducting a systematic examination of relevant scholarly articles, studies, and empirical findings. To what extent has Big Data been used to identify and counteract online threats? Where do we stand right now, and what are the obstacles we must overcome? When it comes to Big Data-driven cybercrime detection, what new trends and technologies are shaping the future? These questions form the backbone of our investigation as we sift through the vast body of literature in search of insights and the pathways that promise a more secure digital future. As we set out on this intellectual adventure, it is crucial that we keep in mind the dynamic and ever-evolving nature of cyber threats and the need for similarly adaptable and innovative countermeasures. This review aims to shed light on the path towards a more secure digital world by illuminating the benefits of combining Big Data analytics with cybersecurity.

2. LITERATURE REVIEW

A literature review is a systematic examination of existing literature on a specific topic or research question. It provides a comprehensive understanding of the current state of knowledge, identifies research gaps, and establishes the context for new studies. Key functions include knowledge synthesis, contextualization, identifying research gaps, methodological insights, theoretical framework, evidence-based decision-making, and proper citation of sources. The process involves selecting a research topic, searching for relevant sources, evaluating sources, synthesizing information, writing the review, citing sources, and revising and updating.

Table 1: Methodology, Data Source and Key Finding

Citation	Title	Methodology	Data Source	Key Findings
[1]	Automated Emerging Cyber Threat Identification and Profiling Based on Natural Language Processing	Natural Language Processing	Social Media Posts	Identified potential cyber threats by analyzing patterns in network traffic and social media data.
[2]	Improving the Performance of Semantic-Based Phishing Detection System Through Ensemble Learning Method	Machine Learning, Anomaly Detection, Feature Selection	Log Files, User Behavior Data	Developed an ML-based approach that effectively detected anomalies in log files and user behavior, improving cybercrime detection accuracy.
[3]	Classification of Malware Using Visualisation of Similarity Matrices	Data Analytics, Visualization, Pattern Recognition	Network Logs, Sensor Data	Explored various big data analytics techniques and visualization methods for cybersecurity, highlighting the importance of pattern recognition in cybercrime detection.
[4]	"Real Time Crime Detection Using Deep Learning Algorithm,"	Deep learning algorithm	Real-time Network Traffic	Demonstrated the feasibility of real-time cybercrime detection by leveraging big data analytics and predictive modeling on live network traffic streams.
[5]	S.P.O.O.F Net: Syntactic Patterns for identification of Ominous Online Factors	Text Mining, Natural Language Processing, Classification	Dark Web Forums, Hacker Chats	Used text mining and NLP techniques to classify and analyze text from hacker forums, aiding in proactive cybercrime detection and threat intelligence.
[6]	Towards Modelling Insiders Behaviour as Rare Behaviour to Detect Malicious RDBMS Access	Behavior Analytics, Entity Linkage, Data Integration	Employee Activity Logs, HR Data	Focused on detecting insider threats through behavior analytics and integrating data from multiple sources for better threat identification.
[7]	Learning Latent Representation for IoT Anomaly Detection	IoT Data Analysis, Anomaly Detection, Time Series	IoT Sensor Data	Addressed cyber threats in IoT-based systems by implementing anomaly detection techniques on time-series IoT sensor data.
[8]	Utilization of Big Data Analytics for Risk Management	AI Algorithms, Threat	Cybersecurity	Proposed an AI-driven approach to fuse data from

		Intelligence, Data Fusion	Reports, News Articles	cybersecurity reports and news articles, providing enriched cyber threat intelligence for better cybercrime detection.
[9]	Scalable Learning Environments for Teaching Cybersecurity Hands-on	Distributed Systems, Data Storage, Real-time Processing	Network Logs, Cloud Infrastructure	Designed a scalable big data platform to handle large volumes of network logs and enable real-time cybersecurity analytics for quicker threat detection.
[10]	Malware Classification using Deep Learning Techniques	Deep Neural Networks, Transfer Learning, Feature Extraction	Malware Samples, Network Traffic	Utilized deep learning techniques for cybercrime classification, achieving high accuracy in identifying malware from network traffic data.

3. CHALLENGES UNEARTHED IN THE FOREGOING LITERATURE EXAMINATION

The previous literature reviews on in said title, provides valuable insights into the numerous research initiatives in the field. However, there are a few potential problems that may need to be addressed:

a. The limited time frame of the review means that cutting-edge methods for detecting cybercrime with big data analytics are not covered. To ensure completeness, more recent studies should be included [11].

b. In favour of machine learning and data analytics, the table above (Table:1) omits deep learning, graph analytics, and ensemble methods. A broader view and multiple approaches are required [13].

c. The use of network traffic and log files as primary data sources raises concerns about the potential for data bias. Researchers could improve the review's credibility by incorporating information from multiple sources, such as social media, the dark web, and the Internet of Things [14].

d. There are no evaluation criteria for the studies presented here. Accurate measurements are necessary for comparing detection methods and identifying knowledge gaps.

e. Inadequate Difficulties and Limitations: Understanding the challenges and limitations of any strategy is critical for determining its viability. The table entries do not elaborate on the difficulties encountered or the restrictions of the proposed solutions [15].

f. It is not clear how the different approaches complement one another from the table because it does not provide a holistic overview of the research that was reviewed.

g. Future research directions aren't discussed. The literature review does not provide the direction needed by scholars and researchers to advance the field of cybercrime detection using big data analytics.

h. Studies with negative findings or those published in less prestigious journals may be overlooked, leading to a potentially misleading sample table [16].

i. Addressing these concerns in the final research review calls for a more extensive and up-to-date literature search, a broader range of methodology and data sources, precise assessment criteria, a comprehensive analysis, and prospective directions. This will improve the review's credibility and usefulness for cybersecurity experts and researchers.

j. In order to better understand how big data analytics can be used to combat cybercrime, the literature review provides a comprehensive overview of the existing research, strategies, and advancements in this area. Everything from academic journals and conference proceedings to public records and company annual reports [17].

k. The importance of big data analytics in detecting, avoiding, and controlling cyber threats and attacks is demonstrated in this research. Many different kinds of data, such as network and system logs, social media data, user behaviour data, and other data sources, are analysed. Insider attacks, distributed denial of service attacks, spear phishing, and other malware are all solvable with the help of big data analytics [18].

l. Some of the central approaches and algorithms are machine learning models, anomaly detection, pattern recognition, and data mining. We also assess the benefits and drawbacks of using big data analytics in cyber security. Data privacy, scalability, and the need for real-time analysis are all examples of such issues.

A better understanding of the state-of-the-art approaches, best practises, and research gaps in the domain of cybercrime detection using big data analytics is provided in this literature review, which is useful for researchers, practitioners, and policymakers working to strengthen cyber defence mechanisms in today's increasingly interconnected digital landscape.

4. STRATEGIES FOR ADDRESSING LIMITATIONS OF PRIOR STUDIES

a. The previous literature review on ‘Artistic rendition’ of "A Literature Review of Big Data for Cybercrime Detection" provides valuable insights into the numerous research initiatives in the field. However, there are a few potential problems that may need to be addressed:
b. The limited time frame of the review means that cutting-edge methods for detecting cybercrime with big data analytics are not covered. To ensure completeness, more recent studies should be included.
c. In favour of machine learning and data analytics, the table above (Table :1) omits deep learning, graph analytics, and ensemble methods. A broader view and multiple approaches are required.
d. The use of network traffic and log files as primary data sources raises concerns about the potential for data bias. Researchers could improve the review's credibility by incorporating information from multiple sources, such as social media, the dark web, and the Internet of Things.
e. There are no evaluation criteria for the studies presented here. Accurate measurements are necessary for comparing detection methods and identifying knowledge gaps.
f. Inadequate Difficulties and Limitations: Understanding the challenges and limitations of any strategy is critical for determining its viability. The table entries do not elaborate on the difficulties encountered or the restrictions of the proposed solutions.
g. It is not clear how the different approaches complement one another from the table because it does not provide a holistic overview of the research that was reviewed.

Future research directions aren't discussed. The literature review does not provide the direction needed by scholars and researchers to advance the field of cybercrime detection using big data analytics.

Studies with negative findings or those published in less prestigious journals may be overlooked, leading to a potentially misleading sample table.

Addressing these concerns in the final research review calls for a more extensive and up-to-date literature search, a broader range of methodology and data sources, precise assessment criteria, a comprehensive analysis, and prospective directions. This will improve the review's credibility and usefulness for cybersecurity experts and researchers.

In order to better understand how big data analytics can be used to combat cybercrime, the literature review " Big Data for Cybercrime Detection " provides a comprehensive overview of the existing research, strategies, and advancements in this area. Everything from academic journals and conference proceedings to public records. The importance of big data analytics in detecting, avoiding, and controlling cyber threats and attacks is demonstrated in this research. Many different kinds of data, such as network and system logs, social media data, user behaviour data, and other data sources, are analysed. Insider attacks, distributed denial of service attacks, spear phishing, and other malware are all solvable with the help of big data analytics.

Some of the central approaches and algorithms are machine learning models, anomaly detection, pattern recognition, and data mining. We also assess the benefits and drawbacks of using big data analytics in cyber security. Data privacy, scalability, and the need for real-time analysis are all examples of such issues.

A better understanding of the state-of-the-art approaches, best practises, and research gaps in the domain of cybercrime detection using big data analytics is provided in this literature review, which is useful for researchers, practitioners, and policymakers working to strengthen cyber defence mechanisms in today's increasingly interconnected digital landscape.

5. CONCLUSION

In conclusion, this literature analysis on ‘Artistic rendition’ of "A Literature Review of Big Data for Cybercrime Detection" provides useful information on its research and progress. Numerous issues must be fixed to improve the review's reliability and practicality. Newer big data analytics-based cybercrime detection trends are not included in the study due to time constraints. More recent studies will help complete the evaluation.

Diversity of Approaches: The example table lacks deep learning, graph analytics, and ensemble approaches, limiting the study. Using more research methods can give a more complete picture of the field.

Data bias can result from overusing network traffic and log files. Include social media, the dark web, and the Internet of Things for a more complete and objective evaluation.

The effectiveness of detection methods is hard to assess without assessment measures. Having clear evaluation criteria simplifies direct comparisons.

Insufficient Challenges and Limitations: Understanding a solution's challenges and constraints is essential to assessing its feasibility. The evaluation must thoroughly examine these factors.

Disjointed Perspective: The table's layout makes it difficult to see how research strategies work together. Comprehensive research is needed to find synergies and research gaps.

Failure to provide future research directions hurts the field. The review should suggest future research and development.

A review should include both positive and negative results to account for publication bias, regardless of paper origin.

The review must conduct a more current literature search, include more methodology and data sources, use trustworthy assessment criteria, critically examine obstacles and constraints, and suggest further research to address these issues. These recommendations will make the literature review credible and useful for practitioners, researchers, and policymakers who want to improve cyber defence systems by analysing large amounts of data. The literature study's many data sources and methods demonstrate big data analytics' role in cybercrime detection. Insider attacks, DDoS, phishing, and malware detection are detailed. We discuss big data analytics' cybersecurity challenges and rate machine learning, anomaly detection, pattern recognition, and data mining models. Academics, industry professionals, and policymakers working to improve cyber defences in the modern, interconnected digital world can use the literature study.

REFERENCES

[1] R. Marinho and R. Holanda, “Automated Emerging Cyber Threat Identification and Profiling Based on Natural Language Processing,” IEEE Access, pp. 1–1, 2023, doi: https://doi.org/10.1109/access.2023.3260020.

[2] A. Maini, N. Kakwani, R. B, S. M K, and B. R, “Improving the Performance of Semantic-Based Phishing Detection System Through Ensemble Learning Method,” IEEE Xplore, Oct. 01, 2021. https://ieeexplore.ieee.org/document/9641614 (accessed Feb. 21, 2023).

[3] S. Venkatraman and Mamoun Alazab, “Classification of Malware Using Visualisation of Similarity Matrices,” Nov. 2017, doi: https://doi.org/10.1109/ccc.2017.11.

[4] P. Sivakumar, Jayabalaguru. V, Ramsugumar. R, and Kalaisriram. S, “Real Time Crime Detection Using Deep Learning Algorithm,” IEEE Xplore, Jul. 01, 2021. https://ieeexplore.ieee.org/document/9526393 (accessed Jun. 24, 2023).

[5] V. R. Mohan, R. Vinayakumar, K. P. Soman, and Prabaharan Poornachandran, “S.P.O.O.F Net: Syntactic Patterns for identification of Ominous Online Factors,” IEEE Symposium on Security and Privacy, May 2018, doi: https://doi.org/10.1109/spw.2018.00041.

[6] Muhammad Imran Khan, B. OrSullivan, and S. N. Foley, “Towards Modelling Insiders Behaviour as Rare Behaviour to Detect Malicious RDBMS Access,” Dec. 2018, doi: https://doi.org/10.1109/bigdata.2018.8622047.

[7] L. Vu, Van Loi Cao, Quang Uy Nguyen, D. N. Nguyen, D. Hoang, and E. Dutkiewicz, “Learning Latent Representation for IoT Anomaly Detection,” IEEE transactions on cybernetics, vol. 52, no. 5, pp. 3769–3782, May 2022, doi: https://doi.org/10.1109/tcyb.2020.3013416.

[8] R. Santhikumar, K Kartillkayani, M. K. Mishra, S. Thota, I. S. Beschi, and B. Mishra, “Utilization of Big Data Analytics for Risk Management,” 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Sep. 2022, doi: https://doi.org/10.1109/icirca54612.2022.9985709.

[9] J. Vykopal, P. Čeleda, P. Seda, V. Švábenský, and D. Tovarňák, “Scalable Learning Environments for Teaching Cybersecurity Hands-on,” IEEE Xplore, Oct. 01, 2021. https://ieeexplore.ieee.org/document/9637180

[10] Bhavya Dawra, A. Chauhan, R. Rani, A. Dev, P. Bansal, and A. Sharma, “Malware Classification using Deep Learning Techniques,” Feb. 2023, doi: https://doi.org/10.1109/delcon57910.2023.10127303.

[11] A. Apurva, P. Ranakoti, S. Yadav, S. Tomer, and N. R. Roy, “Redefining cyber security with big data analytics,” IEEE Xplore, Oct. 01, 2017. https://ieeexplore.ieee.org/document/8284476 (accessed Dec. 13, 2020).

[12] D. B. Rawat, R. Doku, and M. Garuba, “Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security,” IEEE Transactions on Services Computing, vol. 14, no. 6, pp. 1–1, 2019, doi: https://doi.org/10.1109/tsc.2019.2907247.

[13] G. Beaumont, Power BI Machine Learning and OpenAI. Packt Publishing Ltd, 2023.

[14] M. Landauer, Florian Skopik, M. Frank, Wolfgang Hotwagner, M. Wurzenberger, and A. Rauber, “Maintainable Log Datasets for Evaluation of Intrusion Detection Systems,” IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 4, pp. 3466–3482, Jul. 2023, doi: https://doi.org/10.1109/tdsc.2022.3201582.

[15] A. N. Jahromi, H. Karimipour, A. Dehghantanha, and K.-K. R. Choo, “Toward Detection and Attribution of Cyber-Attacks in IoT-enabled Cyber-physical Systems,” IEEE Internet of Things Journal, pp. 1–1, 2021, doi: https://doi.org/10.1109/jiot.2021.3067667.

[16] Nitin Kumar Radke, Shailendra Singh Tomar, and A. Rajan, “Study on Machine Learning Models for IPv6 Address Lookup in Large Block Lists,” Feb. 2023, doi: https://doi.org/10.1109/ncc56989.2023.10068091.

[17] J. Nicholls, A. Kuppa, and N.-A. Le-Khac, “Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape,” IEEE Access, vol. 9, pp. 163965–163986, 2021, doi: https://doi.org/10.1109/access.2021.3134076.

[18] S. Eftimie, R. Moinescu, and C. Racuciu, “Spear-Phishing Susceptibility Stemming From Personality Traits,” IEEE Access, vol. 10, pp. 73548–73561, 2022, doi: https://doi.org/10.1109/access.2022.3190009