ISSN Online: 2821-1936

Transactions on Data Analysis in Social Science

Journal Homepage: https://transoscience.ir

Analysis of Scopus-Indexed Documents on "Malware Detection and Analysis Using Machine Learning and Federated Learning"

A. Hatami[1], P. Hajipour[2,*], H. Eftekhari[3]

[1] MSc Student, Computer Engineering, Computer Networks, Faculty of Electrical and Computer Engineering, Islamic Azad University, North Tehran Branch, Tehran, Iran
[2] Faculty Member, Satellite Communications Group, Communication Technology Research Institute, Research Institute of Communications and Information Technology, Tehran, Iran
[3] CEO, Science and Technology Watch Company, Tehran, Iran

ARTICLE INFO

Article History:
Received 4 January 2025
Received in revised form 22 February 2025
Accepted 22 March 2025
Available online 26 March 2025

Keywords:
Malware, Machine Learning, Federated Learning, Bibliometrics

ABSTRACT

This study provides a comprehensive bibliometric analysis of research trends in malware detection and analysis within communication networks, with particular emphasis on the application of machine learning and federated learning techniques. Using Bibexcel and VOS viewer, a total of 2,915 research documents indexed in the Scopus database between 2008 and 2024 were systematically examined. The analysis explores publication trends, key contributing countries, frequently cited works, and core thematic areas in the field. Statistical findings reveal that concepts such as *malware*, *machine learning*, and *malware detection* dominate scholarly discussions, highlighting their central role in advancing detection frameworks. Moreover, the study identifies India, the United States, and China as the top three leading contributors in terms of research output, reflecting their growing academic and industrial engagement in cybersecurity innovation. Emerging trends such as federated learning indicate a strong research orientation toward privacy-preserving and decentralized approaches, which are becoming increasingly critical in large-scale and distributed communication systems. Overall, the study provides valuable insights into the intellectual structure and global research landscape of malware detection, offering guidance for future studies and the development of more robust, intelligent, and collaborative defense mechanisms against evolving cyber threats.

1. INTRODUCTION

With the advancement of communication networks and their increasing intelligence, many tasks have become more streamlined and efficient. However, hacker groups and cybercriminals exploit various tools and methods, such as developing and deploying malware, to disrupt the functioning of these networks. Consequently, behavioral analysis of malware plays a crucial role in identifying threats and preventing subsequent attacks. Malware analysis

* Corresponding Author: hajipour@itrc.ac.ir
Faculty Member, Satellite Communications Group, Communication Technology Research Institute, Research Institute of Communications and Information Technology, Tehran, Iran

http://dx.doi.org/10.47176/TDASS.2025.46

generally involves two approaches: static and dynamic. In the static mode, the target file is analyzed in terms of content using reverse engineering tools without executing it. In the dynamic mode, the file is executed to examine its behavior under various conditions. Nowadays, cybercriminal groups increasingly adopt a **hybrid analysis** approach, which combines both static and dynamic methods [1].

Currently, activities across various sectors including economic, healthcare, and educational domains are predominantly conducted digitally. Therefore, malware detection and analysis constitute a key component in preventing system breaches and disruptions. Artificial intelligence (AI) and machine learning (ML) algorithms can provide more effective management strategies for combating such malware in communication networks. In this context, companies such as SentinelOne and CrowdStrike are recognized as leading organizations in the field [2]. Among the most significant threats are Advanced Persistent Threats (APTs), which can impose substantial costs on communication networks. To enhance cybersecurity, methods such as Lateral Movement (LM) combined with Remote Desktop Protocol (RDP) have been proposed to achieve successful authentication and prevent malicious host access [3].

The Internet of Things (IoT) is a rapidly expanding technology, creating a high level of connectivity between devices and humans. Consequently, detecting and analyzing threats in IoT environments is critical and necessary. Recent studies have investigated the use of various Federated Learning (FL) architectures as a decentralized approach, where parameters are exchanged without direct sharing, enhancing data privacy. These studies have also suggested security-oriented centralized FL models to meet cybersecurity requirements [4].

Given the importance of malware detection and the need to examine various scientific dimensions for combating potential malware affecting network quality, this study investigates the domain using two complementary approaches: machine learning and federated learning, framed through bibliometric analysis. Bibliometric tools provide a library-based approach, leveraging major databases such as IEEE, WoS, and Scopus, to extract key indicators in any scientific domain, including the number of published documents, leading countries, top authors, and the frequency of keywords in different sections of research articles [5].

The structure of this paper is as follows: Section 2 reviews the research background. Section 3 describes the methodology employed. Section 4 presents the bibliometric-based findings and results. Finally, Section 5 offers the study's conclusions.

2. RESEARCH BACKGROUND

In reference [6], the growth, analysis, and detection of Android-based malware were investigated using a bibliometric approach. This study examined 5,622 documents published between 2010 and 2019 in the Web of Science (WoS) database using a search strategy focused on Android malware, ultimately extracting 1,278 relevant research articles. The results indicate that Asia leads this domain, contributing 40.5% of the publications. The years 2017 and 2016 accounted for the highest output, with 254 and 241 publications, respectively. The fields of Computer Science and Engineering dominate this research area, comprising 86.1% and 38% of the publications, respectively. Among the authors, F. Mercaldo from Italy leads the field with 33 publications.

In reference [7], the study focused on malware and its destructive effects on internet users, also using bibliometric analysis. The research examined documents published between 2005 and 2015 in WoS. The year 2014 recorded the highest number of publications, totaling 242 articles. Regionally, China led in Asia with 268 publications, Germany led in Europe with 102 publications, and the United States topped North America with 679 publications. The Computer Science and Engineering fields accounted for 83.2% and 32.9% of publications, respectively. The most influential journal in this domain was *Lecture Notes in Computer Science*.

Reference [8] addressed the topic of cybersecurity, analyzing its impact across different areas, such as internet-based networks, using bibliometric methods. The study covered publications from 2011 to 2021 in WoS. The results indicated that 2021 had the highest scientific output, with 2,321 published documents. The journal *IEEE ACCESS* led the field with 490 publications. The Chinese Academy of Sciences ranked first among institutions, contributing 203 publications, and the Chinese researcher Y. Wang emerged as the most prolific author with 40 published papers. Globally, the United States, China, and India are the top three countries in this research area.

3. RESEARCH METHODOLOGY

The results presented in this study are based on an analysis of documents including journal articles, conference papers, books, reports, and other publications related to malware analysis and detection using machine learning (ML) and federated learning (FL), employing bibliometric tools. To achieve this, relevant publications in this domain were extracted and systematically analyzed.

In the first step, a comprehensive search was conducted in the Scopus citation database to identify publications in the target research area, following the procedure described below:

("Malware analysis" OR "malware detection") AND ("federated learning" OR "Machine learning")

The number of retrieved documents as of August 25, 2023 amounted to 2,915 records (including articles, books, and other types of publications) from the Scopus database. The relevant details of these documents are reported in the Findings section below.

4. RESEARCH FINDINGS

The distribution of document types published worldwide in this domain such as conference papers, journal articles, books, and book chapters is presented in Figure 1. As illustrated, conference papers, accounting for over 57% of the publications, represent the largest proportion of documents in this research area.

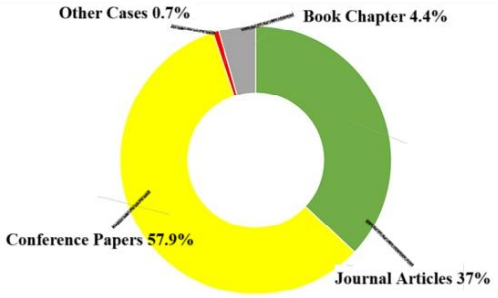

Fig. 1. Types of publications worldwide in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

The total number of publications and their trends in this field are shown in Figure 2. It should be noted that the number of publications for 2023 and 2024 is not yet finalized.

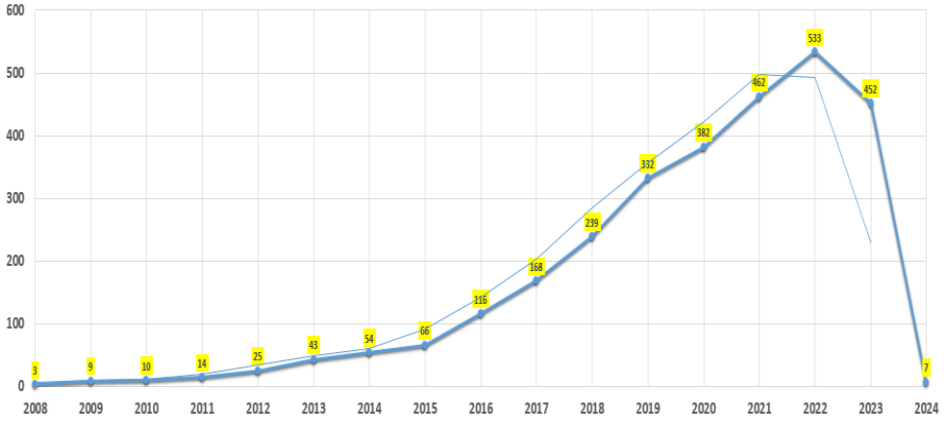

Fig. 2. Number and trend of publications worldwide by year in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

The leading authors in this field, based on the highest number of publications, are shown in Figure 3.

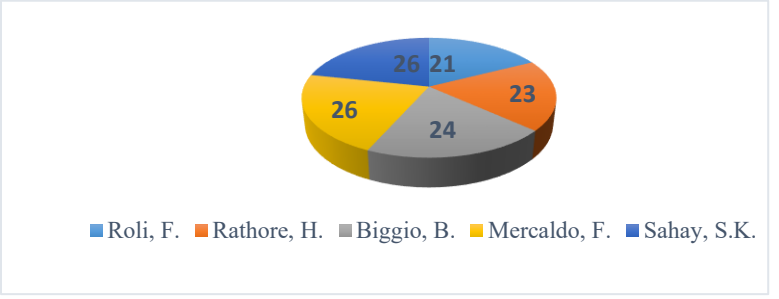

Fig. 3. Leading authors worldwide in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

Additionally, the top five universities, research centers, or companies worldwide with the highest number of publications in this field are presented in Table 1.

Table 1. Leading universities, research centers, and companies worldwide in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

Number of Publications	University / Research Center
43	Chinese Academy of Sciences
36	Birla Institute of Technology and Science, Pilani
33	Beijing University of Posts and Telecommunications
33	Università degli Studi di Cagliari
32	University of Chinese Academy of Sciences

The ranking of countries with the highest number of publications worldwide is also shown in Figure 4. As illustrated, India, the United States, and China occupy the top positions in this ranking.

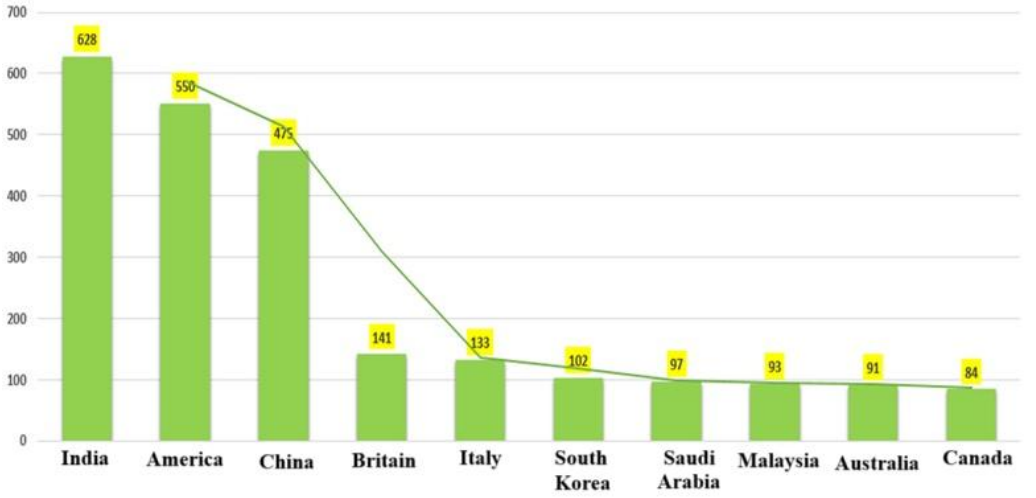

Fig. 4. Countries with the highest number of publications worldwide in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

The distribution and volume of publications across different subject areas are shown in Figure 5. As illustrated, the fields of Computer Science and Engineering account for the largest share of publications in this research domain.

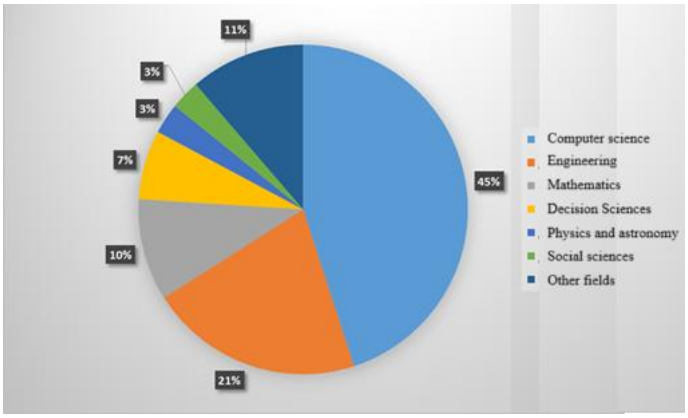

Fig. 5. Distribution and percentage of publications across different subject areas in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

To map the knowledge structure of this research domain worldwide, the publications retrieved from the Scopus database were initially analyzed using Bibexcel software. In this step, the words appearing in the documents were extracted, and the most relevant terms were selected for visualization in the network. This process eliminates less significant terms, allowing the network to focus only on the key terms in this field.

In the next step, the output from Bibexcel was imported into VOSviewer to visualize the knowledge structure of the domain. The first output from this software is shown in Figure 6. The size of the circles in the figure represents the frequency of each term within the publications in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning." As seen in the figure, terms such as "Machine Learning," "Malware Detection," and "Mobile Security" appear in larger circles, indicating they are the most frequently occurring terms in this research area.

In this map, each color represents a distinct thematic cluster. As illustrated, the terms are grouped into seven thematic clusters.

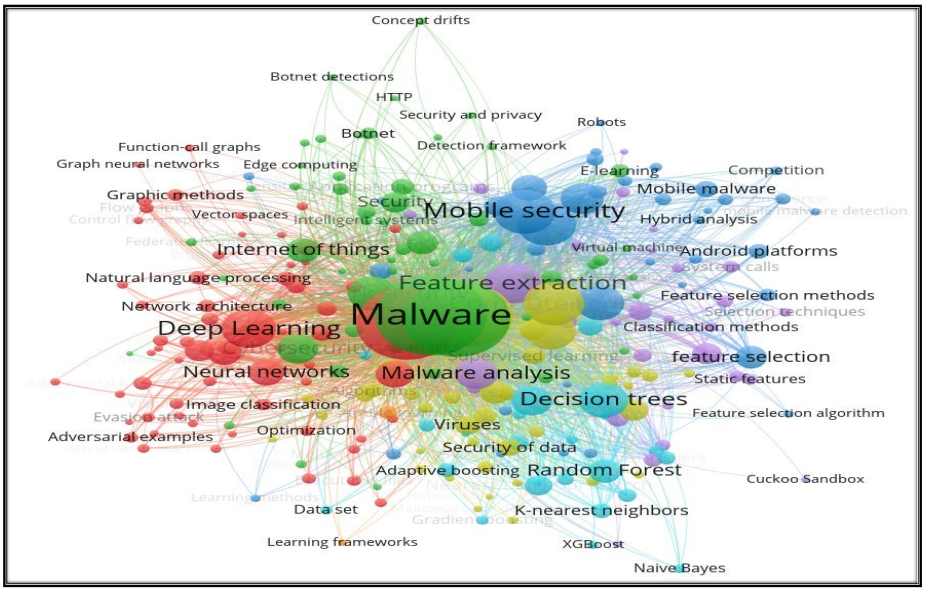

Fig. 6. Co-occurrence network of terms in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

Additionally, the density and distribution of terms in this research domain are shown in Figure 7.

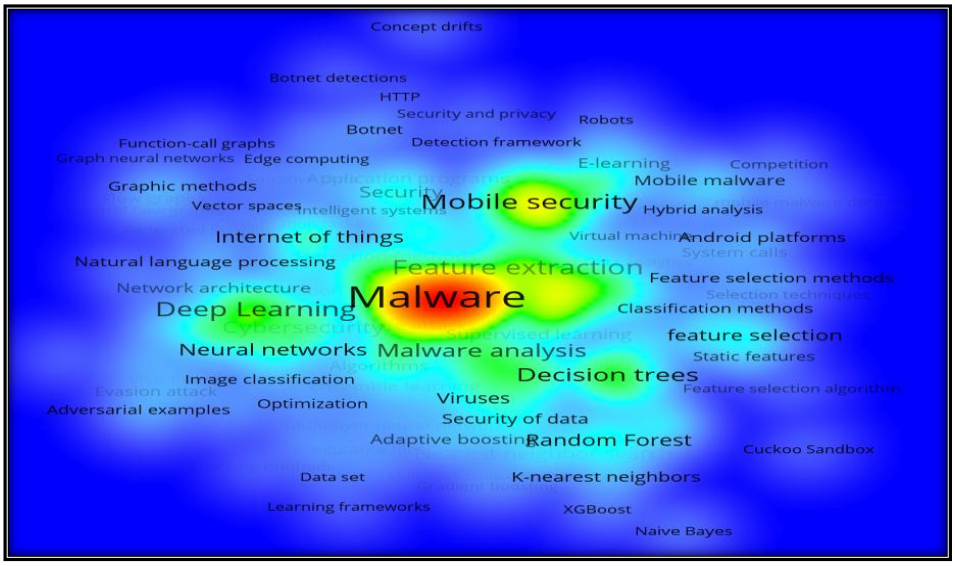

Fig. 7. Density and distribution of terms in the co-occurrence network in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

In Figure 7, the highest term density in the network is indicated in red. Similarly, the yellow, green, and blue areas represent progressively lower densities. The distance between terms also conveys meaningful information: a shorter distance between two terms indicates that they frequently co-occur in the documents, whereas a larger distance suggests that the two terms rarely appear together.

Additionally, the status of international scientific collaborations among countries in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning" is illustrated in Figure 8.

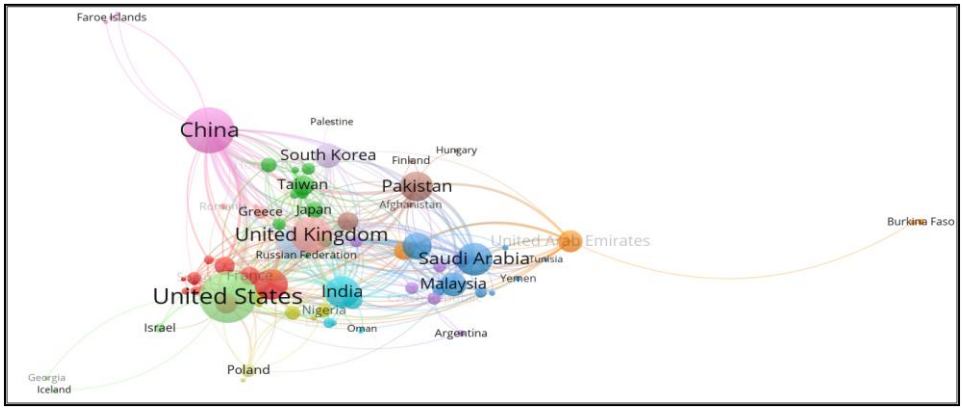

Fig. 8. International scientific collaborations among countries in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

As shown in Figure 8, a total of 82 countries are engaged in international scientific collaborations in this field. Among them, the United States, China, and the United Kingdom have the highest levels of collaboration. The collaboration status of Iran with other countries in this domain is presented in greater detail in Figure 9.

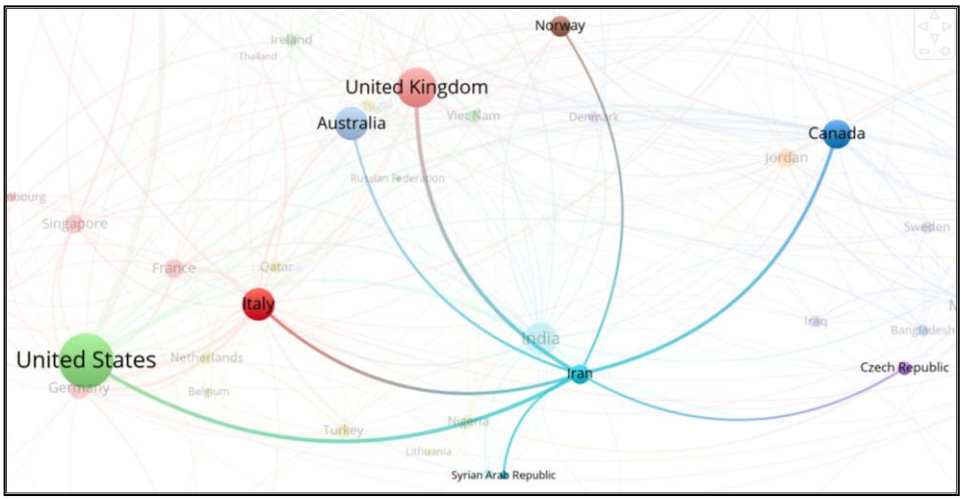

Fig. 9. Collaboration status of Iran in the field of "Malware Analysis and Detection using Machine Learning and Federated Learning."

5. CONCLUSION

Given the critical importance of intelligent methods for combating malware in providing secure services to users, further investigation into this field and the concepts employed in its scholarly outputs is essential. The co-occurrence network analysis of keywords in this domain reveals the presence of seven thematic clusters:

- Cluster 1 consists of 63 keywords, including malware detection, deep learning, and malware classification.
- Cluster 2 contains 55 keywords, such as machine learning, malware, and network security.
- Cluster 3 comprises 40 keywords, including mobile security, Android malware, and static analysis.
- Cluster 4 includes 39 keywords, such as learning systems, cybercrime, and learning algorithms.
- Cluster 5 contains 31 keywords, including feature extraction, malware analysis, and dynamic analysis.
- Cluster 6 comprises 27 keywords, including decision trees, support vector machines, and random forests.
- Cluster 7 includes 3 keywords, such as ensemble learning, detection performance, and learning frameworks.

Additionally, the country-level co-authorship network demonstrates that Iran collaborates scientifically with countries including the United States, Canada, Australia, the United Kingdom, Italy, and others.

Declaration

We acknowledge that we used ChatGPT to enhance the academic writing of our manuscript while ensuring the originality and integrity of our work.

Transparency Statement

The data supporting this study are available upon reasonable request to the corresponding author, subject to ethical and confidentiality considerations.

Acknowledgments

We would like to express our gratitude to all individuals who contributed to this project.

Declaration of Interest

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

REFERENCES

[1] Baker, K. (2023). Malware analysis. CrowdStrike. https://www.crowdstrike.com/cybersecurity-101/malware/malware-analysis

[2] SentinelOne. (n.d.). What is malware detection? Retrieved August 31, 2025, from https://www.sentinelone.com/cybersecurity-101/what-is-malware-detection

[3] Bai, T., Bian, H., Abou Daya, A., Salahuddin, M. A., Limam, N., & Boutaba, R. (2019). A machine learning approach for RDP-based lateral movement detection. In 2019 IEEE 44th Conference on Local Computer Networks (LCN) (pp. 242–245). IEEE. https://doi.org/10.1109/LCN44214.2019.8990853

[4] Venkatasubramanian, M., Lashkari, A. H., & Hakak, S. (2023). IoT malware analysis using federated learning: A comprehensive survey. IEEE Access, 11, 5004–5018. https://doi.org/10.1109/ACCESS.2023.3235389

[5] Xu, B., Li, Y., & Yu, X. (2020). A scientometric analysis of malware detection research based on CiteSpace. In X. Zhang, Y. Li, & F. Chen (Eds.), Machine Learning for Cyber Security (ML4CS 2020): Lecture Notes in Computer Science (Vol. 12486, pp. 100–110). Springer. https://doi.org/10.1007/978-3-030-62223-7_9

[6] Mat, S. R., Ab Razak, M. F., Kahar, M. N., Arif, J. M., Mohamad, S., & Firdaus, A. (2021). Towards a systematic description of the field using bibliometric analysis: Malware evolution. Scientometrics, 126, 2013–2055. https://doi.org/10.1007/s11192-020-03834-6

[7] Ab Razak, M. F., Anuar, N. B., Salleh, R., & Firdaus, A. (2016). The rise of "malware": Bibliometric analysis of malware study. Journal of Network and Computer Applications, 75, 58–76. https://doi.org/10.1016/j.jnca.2016.08.022

[8] Sharma, D., Mittal, R., Sekhar, R., Shah, P., & Renz, M. (2023). A bibliometric analysis of cyber security and cyber forensics research. Results in Control and Optimization, 10, 100204. https://doi.org/10.1016/j.rico.2023.100204