Scientometric Analysis and Knowledge Mapping of Cybersecurity

Fahad Alqurashi, Istiak Ahmad
Department of Computer Science-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah 21589, Saudi Arabia

***Abstract*—Cybersecurity research includes several areas, such as authentication, software and hardware vulnerabilities, and defences against cyberattacks. However, only a limited number of cybersecurity experts have a comprehensive understanding of all aspects of this sector. Hence, it is vital to possess an impartial comprehension of the prevailing patterns in cybersecurity research. Scientometric analysis and knowledge mapping may effectively detect cybersecurity research trends, significant studies, and emerging technologies within this particular context. The main aim of this research is to comprehend the developmental trend of the academic literature about the concepts of "malware detection" and 'cybersecurity'. We collected 9,967 publications from January 2019 to December 2023 and used the Citespace tool for scientometric analysis. This study found six co-citation clusters,namely malware classification, evading malware classifier, android malware detection, IoT network, CNN, and ransomeware families. Additionally, this study discovered that the top contributing countries are the USA, China, and India based on the citation count, and the Chinese Academy of Science, the University of California, and the University of Texas are the top contributing institutions based on the frequency of the publications.**

Keywords*—*Cybersecurity; cyber threats; scientometric analysis; bibliomatic analysis

I. Introduction

The Cybercrime in Australia series [1] is intended to shed light on the victimisation and damages caused by cybercrime among computer users in Australia. The data originates from a survey conducted in early 2023, which included 13,887 individuals who use computers. Before the survey, 27% of participants reported encountering online abuse, 22% encountered malware, 20% faced stolen identities, and 8% fell victim to scams and fraud. Cybersecurity research encompasses a variety of domains, including authentication, software and hardware vulnerabilities, and defences against cyberattacks. However, only a small proportion of cybersecurity professionals have a complete comprehension of all facets of this industry. As a consequence, it is of the utmost importance to develop an objective understanding of the prevalent trends in the field of cybersecurity research. Scientometric analysis and knowledge mapping have the potential to successfully identify patterns, major studies, and new technologies in cybersecurity, transportation [2], health, etc., using different sources such as research articles and newspapers [3], [4]. Researchers, experts, and authorities have to comprehend malware detection technologies and their evolution in cybersecurity. The conceptual structure and dynamic growth of cybersecurity research can be shown by applying these bibliometric approaches to the vast malware detection literature. This method emphasises the most significant contributions and the multidisciplinary links that develop malware detection techniques. Since online hazards have increased dramatically, virus detection technologies are essential for digital security. As malware uses polymorphism and metamorphism to avoid detection, the cybersecurity sector has developed new detection methods. This ongoing arms race between threat actors and defenders requires a thorough examination of research and technological trends. Scientists may categorise malware detection literature by methodology, application fields, and efficacy using scientometric analysis. This study gives a macro picture of the research area, directing future research and technology implementation. Knowledge mapping in malware detection and cybersecurity provides a visual and analytical approach for navigating this field's vast information landscape. It helps explain essential ideas, research links, emerging topics, and technology. Stakeholders may identify prominent research fronts and scientific discourse development using co-citation analysis, co-authorship networks, and keyword co-occurrence mapping. This comprehensive picture helps identify knowledge gaps and encourages collaboration, guiding global cybersecurity efforts towards more robust and adaptable malware detection techniques. This project uses scientometric analysis and knowledge mapping to lay the groundwork for cyber security breakthroughs and safe digital environments for future generations.

a) The Aim and Objectives: The main aim of this research is to comprehend the developmental trend of the academic literature about the concepts of "malware detection" and 'cybersecurity'. This scientometric research aims to analyse the development trend of academic literature specifically focused on "malware detection" and 'cybersecurity'.

- To comprehend the collaboration pattern and analyse the research domain.
- To discover the citation trends from 2019 to 2023.
- To discover the countries, institutions, and keywords involved in the domain of malware detection and cybersecurity.

The remainder of the paper is organised as follows: Section II discusses the similar studies and establishes the research gap. Section III discusses the methodology, including the dataset (Section III-A) and scientometric analysis (Section III-B). Section IV discusses the research outcome. Section V concludes by discussing future work.

II. Literature Review

The quantitative analytics discipline of scientometrics is used to determine and evaluate the volume of research conducted in any given field. In the scientific community, researchers disseminate their findings via a variety of publishing methods. There are several reseach work has been done on scientometric analysis. Raj et al. [5] employed scientometric analysis to discover the knowledge of collaborations, authorship, citations, countrywise, etc. They collected 2720 articles on "cybersecurity" from 2001 to 2018. In another research, [6] focused on Indian authors publications on "cybersecurity" to get knowledge of research trends, collaborating countries, institutions, and top-cited articles. Makawana and Rutvij [7] performed a bibliometric analysis of 149 research articles from 2015 to 2016. Bolbot et al. [8] proposed research direction in maritime cybersecurity by employing meta-analysis (PRISMA) and systematic reviews. The findings demonstrated that Norway, the UK, the USA, and France are the leading nations in maritime cybersecurity. Omote et al. [9] conducted a scientometric analysis using a scienctometric analysis tool named e-CSTI to examine data on science, technology, and innovation in cyber security research. In this research, authors collected data between 2010 and 2019 and discovered that the USA and China emphasise different research areas. In order to provide an in-depth understanding of the present status of medical device cybersecurity research, this study [10] has identified notable authors, organisations, and journal publishers, as well as significant concepts, approaches, and innovations that are often addressed in relation to medical devices. In order to provide an in-depth understanding of the present status of medical device cybersecurity research, this study has identified notable authors, organisations, and journal publishers, as well as significant concepts, approaches, and innovations that are often addressed in relation to medical devices. The study's findings reveal that the most highly contributing country is the USA, and the technology hubs are the UK and India.

The literature review shows that the existing research focused on limited research articles on cybersecurity that were published before 2020. Additionally, there are some works on specific regions or domains. In this study, we collected a total of 9,967 publications from January 2019 to December 2023 and employed scientometric analysis to understand the citation process, calculate the effect of the study, and describe the creation and development of knowledge on a particular research subject.

III. Methodology

A. Dataset

a) Query: The following query is used to collect dataset from WoS: ALL=("malware") OR ALL=("malware detection") OR ALL=("android Malware") OR ALL=("cyber security") OR ALL=("cyber threats") OR ALL=("cyber attacks") OR ALL=(Cyber-Attack) OR ALL=(RANSOMWARE) OR ALL=(CYBERSECURITY).

We apply several filtering approach to get more specific output of searching. For example, this study only select five-years documents (2019 - 2024) and choose document type as proceeding paper and article, that is written in English language. Furthermore, the filtering criteria only include limited WoS categories, such as Computer Science Information Systems, Computer Science Artificial Intelligence, Computer Science Software Engineering, Computer Science Theory Method, Telecommunications, and Computer Science Interdisciplinary Applications, and research areas, for example, Computer Science, and Telecommunications.

TABLE I. Dataset Analyzing Report

WoS Categories		**Document Types**	
Computer Science Information Systems	5,790	Proceeding paper	5,907
Computer Science Theory Method	4,823	Article	4,118
Computer Science Artificial Intelligence	2,162	**Countries**	
Computer Science Software Engineering	1,852	USA	2,458
Telecommunications	1,816	China	1,668
Computer Science Interdisciplinary App.	1,575	India	828
Research Areas		England	695
Computer Science	9,448	Germany	540
Telecommunications	1,816	Australia	506

This study collected a total of 9,967 Publications, where those publications 62,377 times total cited and 52,546 times without self-citation. The total citing articles are 37,384 and without self-citation are 33,747 with H-Index equal to 82.

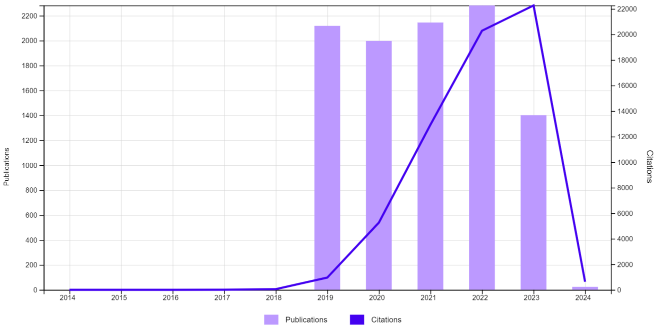

Fig. 1. Articles citation report generated from WoS.

Fig. 1 depicts the citation report of the collected 5-years dataset from Web of Science. The details of the dataset is listed in Table I.

B. Scientometric Analysis

The quantitative investigation of scientific research is referred to as scientometrics. Using extensive datasets of research publications, it allows for the understanding of the citation process, calculates the effect of the study, and describes the creation and development of knowledge on a particular research subject. While it is still possible to miss literary concepts in traditional investigations, scientometric approaches allow academics to analyse a significant quantity of bibliometric data and identify systematic conclusions connected to literature. This investigation employed CiteSpace [11], a Java-based programme that analyses and visualises co-citation networks, for scientometric analysis. The purpose of the tool is to pinpoint turning points and new trends in a certain field. It provides unique benefits for presenting and evaluating scientific data to enable more accurate interpretation of earlier research by painstakingly creating a multitude of easily understood visualisations that may help reveal the implications hidden in a vast body of knowledge. Some importance terms used in scientometric analysis are co-citation analysis, Burst Strength, Burst Begin-End, Degree, Centrality and Sigma.

In the network generated by CiteSpace, two quantifiable markers may be used to identify important nodes: the burst strength and betweenness centrality. The proportion of the shortest path between two clusters to the total of these shortest routes is used to calculate node betweenness centrality.

$$Centrality(node_x) = \sum_{x \neq y \neq z} \frac{\gamma_{yz}(x)}{\gamma_{yz}} \quad (1)$$

In Eq. 1, $\gamma_{yz}(x)$ represents the count of pathways that go via node x, whereas γ_{yz} represents the count of the shortest routes linking node y and node z. The burst identification technique was used to identify sudden fluctuations in citations at certain time periods. The process of calculating citation bursting strength begins with acquiring and importing pertinent bibliometric information, then implements Kleinberg's method to analyse the citation timeline for every document inside the collection. Citation burst strength is determined by statistically evaluating the increase in citation frequency within a certain time period in comparison to periods with no significant increase. An article with a significant burst strength demonstrates a notable rise in its citation rate, indicating an enhanced level of impact or significance during the burst timeframe. The citation degree is calculated by calculating the number of linkages a node has with adjacent nodes in the network. A larger citation degree shows more direct citations, signifying more impact or significance on the subject. Conversely, Sigma is computed by multiplying the number of citations by betweenness centrality, which expresses the frequency at which a node acts as a link between other nodes. Papers with high sigma values are often both highly cited and influential in bridging various disciplines or concepts within the academic field.

In addition, CiteSpace provides scientometric analysis that includes an investigation of countries, organisations, and the co-occurrence of keywords.

IV. Results and Discussion

Cluster analysis is a prevalent approach to finding hidden contextual patterns in knowledge discovery. Through the use of cluster analysis, an extensive repository of data from research is divided into discrete units according to the relative strength of word correlation. This facilitates the identification of research themes, patterns, and their connections within a certain field of study. In this study, six co-citation clusters were identified using the log-likelihood ratio (LLR) technique. This was possible since the clusters created by LLR had excellent quality, with high intra-class and low inter-class similarity. Additionally, based on the uniqueness and coverage of each cluster, LLR chooses a label based on the keywords of the texts cited in each cluster. Cluster labelling quality is determined by the variety, depth, and breadth of terms formed from keywords in articles. The label supplied for each cluster identifies the focus of that cluster. Fig. 2 shows the cluster analysis using co-citation analysis, demonstrating the timeline of each cluster. We discovered six clusters including malware classification (ClusterID=0), evading malware classifier (clusterID=1), android malware detection (clusterID=2), iot network (clusterID=3), convolutional neural network (clusterID=4), and ransomware families (clusterID=5).

A. *Cluster Analysis*

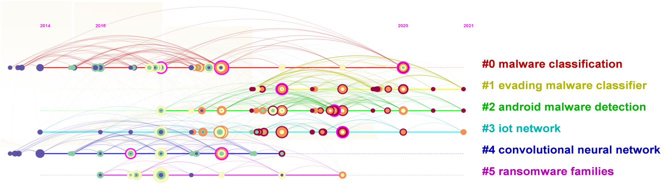

Fig. 2. Cluster analysis.

Table II shows the cluster network summary by listing the top 20 research publications sorted by burst strength. The list includes all details, such as, publication year, burst strength, burst begin-end, degree, centrality, sigma, frequency, and cluster ID for each publication.

TABLE II. Cluster Network Summary

Ref.	Pub. Year	Burst Strength	Burst Begin-End	Degree	Cent.	Sigma	Freq.	CID
[12]	2015	13.02	2019 - 2020	19	0.06	2.04	42	0
[13]	2015	11.45	2019 - 2020	13	0.03	1.4	37	0
[14]	2020	9.75	2021 - 2023	17	0.02	1.17	42	2
[15]	2018	8.57	2021 - 2023	3	0	1.01	94	3
[16]	2017	8.29	2020 - 2021	12	0.03	1.32	33	0
[17]	2016	8.04	2020 - 2021	2	0.01	1.04	32	5
[18]	2015	7.7	2019 - 2020	4	0	1	25	5
[19]	2019	7.64	2021 - 2023	10	0.06	1.51	33	5
[20]	2017	7.39	2019 - 2020	11	0.01	1.07	24	4
[21]	2015	7.39	2019 - 2020	9	0.01	1.07	24	4
[22]	2016	7.31	2019 - 2021	6	0.02	1.16	121	5
[23]	2019	6.68	2021 - 2023	24	0.06	1.52	74	2
[24]	2018	4.93	2020 - 2021	21	0.06	1.3	53	2
[25]	2019	4.8	2021 - 2023	23	0.11	1.65	57	2
[26]	2016	4.79	2019 - 2021	9	0.08	1.47	80	3
[27]	2019	4.76	2021 - 2023	21	0.09	1.48	58	1
[28]	2020	4.6	2021 - 2023	9	0.01	1.07	56	1
[29]	2019	4.51	2021 - 2023	18	0.02	1.1	55	2
[30]	2016	3.93	2019 - 2020	15	0.06	1.24	49	0
[31]	2016	3.21	2020 - 2021	8	0.04	1.14	49	5

1) *Malware classification and evading malware classifier:*

a) Malware Variations: Malware can be classified into several types, including worms, spyware, viruses, trojans, bots, rootkits, ransomware, scareware, and so on.

Worms use software and operating system flaws to propagate to other machines. They do not need to connect to a programme like viruses. Worms may overburden web servers, steal data, install backdoors, and more. Worms' speed and autonomy make them hazardous, causing internet interruptions and severe financial harm to afflicted organisations and individuals. Spyware secretly tracks users' internet activities, keystrokes (keyloggers), and financial data. It may be installed without the user's knowledge via free software downloads or malicious websites. Identity theft and unauthorised access to personal and financial data may result from spyware, which slows system performance and internet connections.

When run, viruses change other computer programmes and implant their own code. Infected systems may malfunction, lose data, and operate poorly. Email attachments, compromised software programmes, and file downloads distribute viruses, which need human input to activate their destructive activities.

Trojans, or Trojan horses, deceive users. They frequently seem like respectable applications but do bad things when run. Trojans, unlike viruses and worms, do not multiply but may provide paths for other malware to steal data or create a zombie machine under an attacker's control. Trojans may gain unauthorised access to systems, stealing data, compromising privacy, and installing further software.

Computer programmes called bots automate jobs. However, malicious bots are used to take control of a computer and employ it in a botnet. DDoS assaults, spamming, phishing, and cryptocurrency mining are all possible with botnets. Botnet machines may be located worldwide, making attacks hard to track. Rootkits stealthily obtain root or administrative access to a computer without users or security software noticing. Rootkits may intercept and modify system operations to mask their presence and other malicious actions, making detection and removal difficult. Rootkits let attackers steal data, monitor user activities, and remotely control a machine.

Ransomware encrypts a victim's data or locks them out of their machine and demands a fee to decode or unlock. Phishing emails, fraudulent ads, and software weaknesses disseminate it. Ransomware may cause considerable data loss, financial harm, and operational interruption until the ransom is paid or files are recovered from backups. Scareware tricks users into thinking their computer has a virus or other major problem to get them to buy needless or hazardous software. It usually appears as pop-up advertising or antivirus software-like security notifications. Scareware may cost money and install spyware or other malware if the user buys it or removes the phoney risks it claims to have found.

b) Types of Malware Detection: The top cited papers for malware detection and classification are [32], [27], [33], [34], [28].

Signature-based Detection: This approach is one of the simplest and traditional techniques for detecting malware. Antivirus software conducts scans on files, executable programmes, and system locations, and then compares them with a database in order to identify any matches. The signature-based approach detects distinct character sequences inside the binary code. Each time a novel kind of malware is released, anti-malware companies must acquire a sample of the new virus, scrutinise it, generate fresh signatures, and distribute them to their customers. Conventionally, domain experts are responsible for manually creating, updating, and distributing the signature bases. This technique is often recognised as being time-consuming and requiring a significant amount of labour. This kind of detection strategy reduces the responsiveness of anti-malware software programmes to emerging threats. It has the potential to enable some malware samples to evade detection and remain undiscovered for an extended time.

Heuristic-based Detection: This approach uses algorithms to analyse the behaviour and features of programmes to discover suspected malware based on abnormal patterns or behaviours. This approach goes beyond signature matching; instead, it examines the code's structure for any unusual traits that might point to a danger, including the inclusion of code that is often used to take advantage of vulnerabilities. By concentrating on characteristics shared by malicious software, heuristic detection may detect newly created or altered malware, although it may produce more false positives than signature-based detection.

Behavior-Based Detection: This approach keeps a check on how software behaves naturally inside the system, rather than employing malware fingerprints to identify threats ahead of time. This method monitors how an application accesses network resources, user data, system files, and processes and checks for malicious activity such as unapproved changes, eavesdropping, or data exfiltration. It can successfully detect polymorphic and previously undiscovered malware that would elude signature-based techniques since it analyses behaviours in real-time. Its emphasis on behaviour, meanwhile, may result in false alerts should benign programmes exhibit anomalous behaviour.

Anomaly-Based Detection: Providing a baseline of typical network or system activity, anomaly-based detection then keeps monitoring for variations from this baseline. Significant discrepancies might be a sign that malware is present. This technique is very helpful in detecting complex assaults and zero-day threats, but it may produce incorrect results if the baseline is not well established.

Sandbox Detection: Malicious programs are run in a virtual environment called a "sandbox" that is isolated from the primary system in sandbox detection. This keeps the system safe while enabling the programme to execute and display its behaviour. Sandboxing works well against malware that may avoid identification by detecting it during analysis or by postponing execution.

Cloud-based Detection: The process of detecting malware now follows a client-server approach using a cloud-based architecture. This involves preventing the execution of unauthorised software programmes listed in a blacklist and verifying the legitimacy of software programmes listed in a whitelist at the user's end. Additionally, any unknown files are analysed at the server side and the results are promptly communicated to the clients. The grey list comprises unfamiliar software files, which may be either harmless or dangerous. Historically, the grey list was either rejected or verified manually by experts in malware analysis. Due to advancements in malware authoring and creation methods, the quantity of file samples on the grey list is consistently growing. As an example, the grey list produced by either Kingsoft or Comodo Cloud Security centre often includes over 500,000 file samples on a daily basis [Ye 2010]. Therefore, it is essential to create intelligent methods to enhance the efficiency and effectiveness of malware detection on the server side of the cloud.

Hybrid Detection Methods: Hybrid methodologies integrate many detection methods to enhance the overall effectiveness of malware detection and minimise the occurrence of false positives. For instance, antivirus software may use a combination of signature-based and behavior-based detection methods to provide extensive safeguarding against both recognised and unrecognised hazards.

Feature Analysis: Static analysis examines PE files without running them. Static analysis targets binary or source codes. If a PE file is compressed using third-party tools like UPX or ASPack Shell, it must be decompressed first. To decompile Windows executables, employ disassembler and memory dumper tools. Memory dumper tools extract protected main memory codes and save them to a file. A memory dump is important for examining packed executables that are hard to deconstruct. Unpacking and decrypting the executable reveals static analysis patterns such Windows API calls, byte n-grams, strings, opcodes, and control flow graphs. Feature extraction is achieved via dynamic analysis methods, such as profiling and debugging, by analyzing the PE files being executed (on a physical or virtual CPU). To do dynamic analysis, a variety of methods may be used, including function parameter analysis, function call monitoring, information flow tracking, and instruction traces.

2) Android malware detection and convolutional neural network: The rapid proliferation of Android malware, its ability to evade detection, and the possible loss of enormous amounts of data assets held on Android devices make Android malware detection and categorization an issue involving big data. Applying deep learning to Android malware detection appears to be a logical and intuitive decision. Nevertheless, scholars and practitioners encounter several obstacles, including the selection of a deep learning architecture, the extraction of features, the evaluation of efficacy, and the acquisition of sufficient high-quality data. This research discovered the top cited papers for android malware detection based on the Co-citation network are [35], [36], multimodal technique [23], significant permission identification [37], intrusion detection dataset [15], Google Playstore Android dataset named AndroZoo [38], and so on.

Fully Connected Network (FCN) [39] has been used in several Android malware detection approaches. The FCN analyzed the AndroidManifest.xml and classes.dex files to extract information such as needed permissions, contextual details, and API calls, which were then used to characterize the Android programs. The activation function employed in the hidden layers is the Parametric Rectified Linear Unit function, as it is very efficient and allows for dynamic modification. In the output layer, Softmax is employed as an activation function.

Convolutional Neural Network (CNN) [36] is employed to identify Android malware in raw opcode sequences. First, the application for Android was disassembled, and its opcode sequences were retrieved for analysis. An opcode embedding layer received one-hot vectors of opcode instructions. The embedding layer enabled the CNN network to gather opcode semantics. Abstract characteristics were extracted using convolution layers. Moreover, a max-pooling layer after each convolution layer selected the most appropriate malware-detecting opcode sequence. The app's maliciousness was determined via a fully linked hidden layer before the output layer. Through cooperative training, the CNN network learned malware patterns from raw opcode sequences without employing any handcrafted features.

Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) [40] can acquire semantic knowledge and connections within sequential data, enabling them to process sequential opcode or bytecode. Xin et al. [41] presented DroidDeep, a DBN-based tool for Android malware detection. It uses approximately 32,000 layers AndroidManifest.xml and classes.dex features. These features include app permissions, activities, components, permissions used, and requests to sensitive APIs. DroidDeep prepares string properties for processing by one-hot encoding them as numerical vectors. The DroidDeep DBN architecture uses unsupervised pre-training to find high-level feature representations and supervised fine-tuning via back-propagation to improve detection. These learnt characteristics are used to train an SVM classifier to detect malware. DroidDeep excels in malware detection with 99.4% accuracy, making it ideal for real-world applications. The stacked Auto-Encoder (AE) in Deep4MalDroid [42] analysed the graph-based characteristics to identify the Android malware.

3) IoT Network: The top cited research papers discovered by this study are [43] and [44]. The first paper discussed Advanced Persistent Threats (APT) detection-related challenges and unsolved issues using ML. Hackers target Internet of Things (IoT) systems for a variety of reasons, including disclosing, shifting, disabling, copying, or obtaining unauthorised access to or using an asset without authorization. The second paper discussed DDoS in the IoT. A Denial-of-Service (DoS) attack is one instance when an attacker uses an authorised host network to transmit a large number of packets to the victim to overwhelm them with messages. On the other hand, port scanning assaults take place when a hacker finds an open port that might be used to launch an attack. As a result, hackers are able to get comprehensive information about the network, such as MAC and Internet Protocol (IP) addresses. The most used datasets for IoT-based threat detection are N-BaIoT [45], Bot-IoT [46], ToN-IoT [47], and Edge-IIoTset [48].

4) Ransomware families: Ransomware is a kind of malware that is used as a means of extortion. Ransomware is a kind of malicious software that covertly infiltrates a victim's system and promptly demands payment in exchange for decrypting the encrypted data [31], [49]. The majority of ransomware families exhibit the following features: device lockout, data deletion and stealing, encryption, and delivering alarming notifications. Ransomware families include Cryptolocker, CryptoWall, CTB-Locker, CrypVault, CoinVault, Filecoder, TeslaCrypt, Tox crypto, VirLock, Reveton, Tobfy, and Urausy.

B. Country Analysis

Fig. 3 shows the node-line country network, in which each node is a country and the line indicates the cooperative links between nations. The amount of articles determines the node's size of the country.

Table III shows the country network summary by listing the top 10 countries sorted by four categories: citation count, degree, centrality, and sigma. The highest-rated countries based on citation counts are the USA (2019), China (1417), India (694), England (523), Australia (416), and so on. Based on degrees, the top countries are England (39), the USA (30), the Netherlands (NL) (29), Belgium (28), France (27), and so on. England and Wales are the highest-rated countries based on centrality and sigma, respectively.

C. Institution Analysis

Fig. 4 shows the node-line institution network, in which each node is a institution and the line indicates the cooperative links between institutions. The amount of articles determines the node's size of the institution.

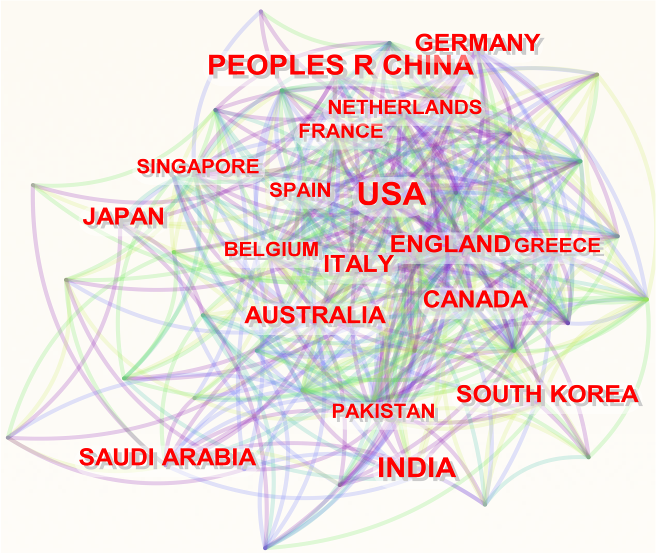

Fig. 3. Country analysis.

TABLE III. Country Network Summary

Citation Count		Degree		Centrality		Sigma	
USA	1979	England	39	England	0.16	Wales	1.01
China	1417	USA	30	USA	0.08	USA	1.00
India	694	NL	29	France	0.08	England	1.00
England	523	Belgium	28	NL	0.07	France	1.00
Australia	416	France	27	Australia	0.07	Australia	1.00
Germany	416	Sweden	26	Italy	0.06	NL	1.00
Italy	363	Italy	25	UAE	0.06	Italy	1.00
South Korea	285	Spain	25	Singapore	0.06	UAE	1.00
Saudi Arabia	269	Pakistan	25	Belgium	0.05	Singapore	1.00
Canada	268	UAE	25	Spain	0.05	Belgium	1.00

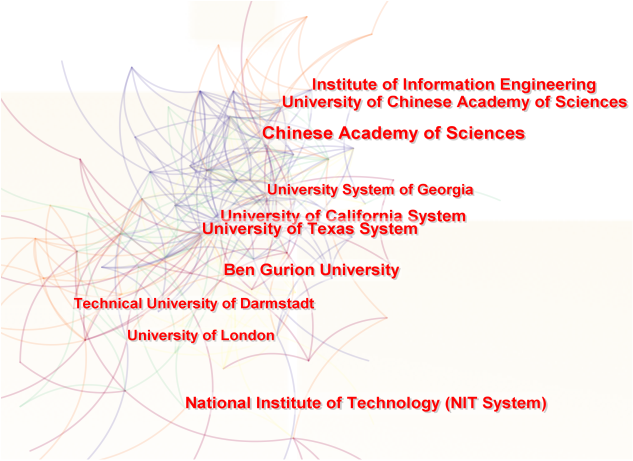
Institute of Information Engineering
University of Chinese Academy of Sciences
Chinese Academy of Sciences
University System of Georgia
University of California System
University of Texas System
Ben Gurion University
Technical University of Darmstadt
University of London
National Institute of Technology (NIT System)

Fig. 4. Institution analysis.

Table IV shows the institution network summary by listing the top 10 institutions sorted by two categories: frequency, and burst-strength. The top-rated institutions based on frequency are Chinese Academy of Sciences (194), University of California (147), University of Texas (127), and so on. Based on burst strength, the highest institutions are University of California Berkeley (11.14), Fraunhofer Gesellschaft (7.63), IMT - Institut Mines-Telecom (7.04), University of North Carolina (5.76), KU Leuven (5.21), and so on.

TABLE IV. Top Institution Network Summary

Frequency-based		Burst Strength-based	
Chinese Academy of Sciences	194	University of California Berkeley	11. 14
University of California	147	Fraunhofer Gesellschaft	7.63
University of Chinese Academy of Sciences	130	IMT - Institut Mines-Telecom	7.04
University of Texas	127	University of North Carolina	5.76
State University of Florida	124	KU Leuven	5.21
Institute of Information Engineering	109	IIT	4.64
Ben Gurion University	109	Texas A&M University	4.35
National Institute of Technology	104	University of Illinois	2.91
University of Georgia	86	University of Illinois Urbana-Champaign	1.41
Nanyang Technological University	84	National University of Singapore	0.97

D. Keywords Analysis

Fig. 5 depicts the keyword co-occurrence network. Keywords are concise and indicative synopses of the content of research studies. Keyword co-occurrence networks may be used to identify the current most prevalent topics in the area of knowledge during a certain time period. The node's size is determined by how often it uses the keywords.

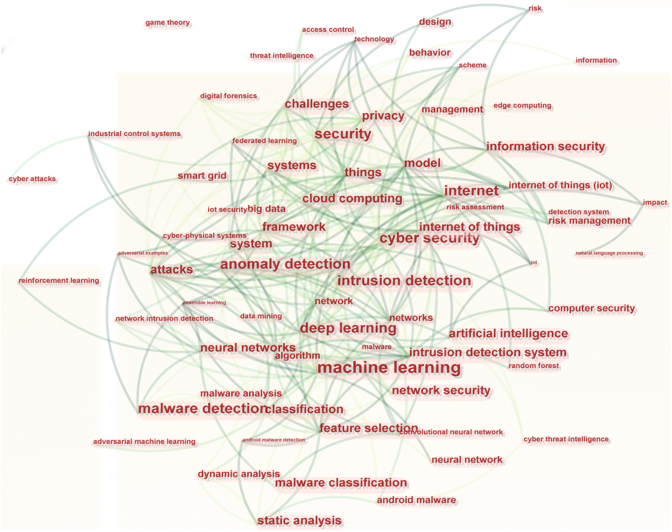

Fig. 5. Keywords analysis.

Table V shows the most frequent terms with frequency from 2019 to 2023. some keywords, such as, federated learning (16), risk (15), ensemble learning (12), adversial examples (12), iot (11), and NLP (11) are mostly used in 2023. The top keywords in 2022 are desgn (40), cyber-physical systems (33), CNN (24), reinforecemnt learning (22), scheme (19), and random forest (18). In 2029, the most frequent used keywords are intrusion detection (369), machine learning (852), deep learning (476), information security (131), cloud computing (111), feature selection (121), static analysis (120), dynamic analysis (60), android malware (48), and data mining (15).

TABLE V. KEYWORD NETWORK SUMMARY

Year	Keywords with Frequency
2023	federated learning (16), risk (15), ensemble learning (12), adversial examples (12), iot (11), NLP (11)
2022	desgn (40), cyber-physical systems (33), CNN (24), reinforecemnt learning (22), scheme (19), random forest (18)
2021	network (74), algorithm (63), risk management (34), industrial control system (26), network intrusion detection (18), adversial machine learning (15)
2020	internet of things (99), digital forensics (15), behavior (39), malware (18), computer security (35), cyber threat intelligence (19), information (16)
2019	intrusion detection (369), machine learning (852), deep learning (476), information security (131), cloud computing (111), feature selection (121), static analysis (120), dynamic analysis (60), android malware (48), data mining (15)

V. CONCLUSION

Cybersecurity is a crucial study issue that is garnering significant attention across all sectors. Mapping cybersecurity research is crucial to assess the level of preparation in cybersecurity skills and identify areas that need improvement. This study aims to discover research needs and peaks in the fields of cyber security, malware detection, and android malware detection. This study performs a scientometric analysis and knowledge mapping of cybersecurity-related papers that were published over the last five years. We collected 9.967 research articles from WoS (see Section III-A). After that, scientometric analysis is performed to analyse research domain patterns, related research knowledge, which is referred to as clusters in this study, and keywords, and finally, discover the most contributing countries and institutions. This study found six clusters: cluster ID=0 for malware classification; cluster ID=1 for evading malware classifier; cluster ID=2 for android malware detection; cluster ID=3 for iot networks; cluster ID=4 for convolutional neural networks; and cluster ID=5 for ransomware families. The United States, China, and India are the top three contributors in terms of citation count. The Chinese Academy of Science, the University of California, and the University of Texas are the top contributing institutions over the last five years. Future work may include analysing the complete literature and comparing the findings to those from the top-ranked journals.

ACKNOWLEDGMENT

The authors acknowledge with thanks the technical and financial support from the Deanship of Scientific Research (DSR) at the King Abdulaziz University (KAU), Jeddah, Saudi Arabia, under Grant No. (IFPIP: 1353-611-1443).

REFERENCES

[1] I. Voce and A. Morgan, *Cybercrime in Australia 2023*. Australian Institute of Criminology, 2023.

[2] I. Ahmad, F. Alqurashi, E. Abozinadah, and R. Mehmood, “Deep journalism and deepjournal v1. 0: a data-driven deep learning approach to discover parameters for transportation,” *Sustainability*, vol. 14, no. 9, p. 5711, 2022.

[3] I. Ahmad, F. AlQurashi, and R. Mehmood, “Machine and deep learning methods with manual and automatic labelling for news classification in bangla language,” *arXiv preprint arXiv:2210.10903*, 2022.

[4] ——, “Potrika: Raw and balanced newspaper datasets in the bangla language with eight topics and five attributes,” *arXiv preprint arXiv:2210.09389*, 2022.

[5] S. Rai, K. Singh, and A. K. Varma, “Global research trend on cyber security: A scientometric analysis,” *Library Philosophy and Practice (e-journal)*, vol. 3339, 2019.

[6] B. Elango, S. Matilda, M. Martina Jose Mary, and M. Arul Pugazhendhi, “Mapping the cybersecurity research: A scientometric analysis of indian publications,” *Journal of Computer Information Systems*, vol. 63, no. 2, pp. 293–309, 2023.

[7] P. R. Makawana and R. H. Jhaveri, “A bibliometric analysis of recent research on machine learning for cyber security,” *Intelligent Communication and Computational Technologies: Proceedings of Internet of Things for Technological Development, IoT4TD 2017*, pp. 213–226, 2018.

[8] V. Bolbot, K. Kulkarni, P. Brunou, O. V. Banda, and M. Musharraf, “Developments and research directions in maritime cybersecurity: A systematic literature review and bibliometric analysis,” *International Journal of Critical Infrastructure Protection*, p. 100571, 2022.

[9] K. Omote, Y. Inoue, Y. Terada, N. Shichijo, and T. Shirai, “A scientometrics analysis of cybersecurity using e-csti,” *IEEE Access*, pp. 1–1, 2024.

[10] O. A. Alfahad, T. Ur Rehman, A. Woodman, E. A. Malaekah, and M. Rasheed, “Mapping knowledge and themes trends in the cybersecurity of medical devices: A bibliometric investigation,” *Science & Technology Libraries*, pp. 1–11, 2023.

[11] C. Chen, “Searching for intellectual turning points: Progressive knowledge domain visualization,” *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl_1, pp. 5303–5310, 2004.

[12] J. Saxe and K. Berlin, “Deep neural network based malware detection using two dimensional binary program features,” in *2015 10th International Conference on Malicious and Unwanted Software (MALWARE)*, 2015, pp. 11–20.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[14] M. K. Alzaylaee, S. Y. Yerima, and S. Sezer, “Dl-droid: Deep learning based android malware detection using real devices,” *Computers & Security*, vol. 89, p. 101663, 2020.

[15] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization.” *ICISSp*, vol. 1, pp. 108–116, 2018.

[16] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, “Adversarial examples for malware detection,” in *Computer Security–ESORICS 2017: 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II 22*. Springer, 2017, pp. 62–79.

[17] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[18] A. Kharraz, W. Robertson, D. Balzarotti, L. Bilge, and E. Kirda, “Cutting the gordian knot: A look under the hood of ransomware attacks,” in *Detection of Intrusions and Malware, and Vulnerability Assessment: 12th International Conference, DIMVA 2015, Milan, Italy, July 9-10, 2015, Proceedings 12*. Springer, 2015, pp. 3–24.

[19] O. Or-Meir, N. Nissim, Y. Elovici, and L. Rokach, “Dynamic malware analysis in the modern era—a state of the art survey,” *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–48, 2019.

[20] K. Tam, A. Feizollah, N. B. Anuar, R. Salleh, and L. Cavallaro, “The evolution of android malware and android analysis techniques,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, pp. 1–41, 2017.

[21] P. Faruki, A. Bharmal, V. Laxmi, V. Ganmoor, M. S. Gaur, M. Conti, and M. Rajarajan, “Android security: A survey of issues, malware penetration, and defenses,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 998–1022, 2015.

[22] T. Klikauer, “Reflections on phishing for phools: The economics of manipulation and deception,” *TripleC*, pp. 260–264, 2016.

[23] T. Kim, B. Kang, M. Rho, S. Sezer, and E. G. Im, “A multimodal deep learning method for android malware detection using various features,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 773–788, 2019.

[24] E. B. Karbab, M. Debbabi, A. Derhab, and D. Mouheb, “Maldozer: Automatic framework for android malware detection using deep learning,” *Digital Investigation*, vol. 24, pp. S48–S59, 2018.

[25] F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, "{TESSERACT}: Eliminating experimental bias in malware classification across space and time," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 729–746.

[26] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[27] D. Ucci, L. Aniello, and R. Baldoni, "Survey of machine learning techniques for malware analysis," *Computers & Security*, vol. 81, pp. 123–147, 2019.

[28] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *Journal of Network and Computer Applications*, vol. 153, p. 102526, 2020.

[29] L. Onwuzurike, E. Mariconti, P. Andriotis, E. D. Cristofaro, G. Ross, and G. Stringhini, "Mamadroid: Detecting android malware by building markov chains of behavioral models (extended version)," *ACM Transactions on Privacy and Security (TOPS)*, vol. 22, no. 2, pp. 1–34, 2019.

[30] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387.

[31] A. Kharaz, S. Arshad, C. Mulliner, W. Robertson, and E. Kirda, "{UNVEIL}: A {Large-Scale}, automated approach to detecting ransomware," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 757–772.

[32] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–40, 2017.

[33] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, and Q. Yu, "A survey of adversarial attack and defense methods for malware classification in cyber security," *IEEE Communications Surveys & Tutorials*, 2022.

[34] Z. Cui, F. Xue, X. Cai, Y. Cao, G.-g. Wang, and J. Chen, "Detection of malicious code variants based on deep learning," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3187–3196, 2018.

[35] J. Qiu, J. Zhang, W. Luo, L. Pan, S. Nepal, and Y. Xiang, "A survey of android malware detection with deep neural models," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–36, 2020.

[36] N. McLaughlin, J. Martinez del Rincon, B. Kang, S. Yerima, P. Miller, S. Sezer, Y. Safaei, E. Trickel, Z. Zhao, A. Doupé *et al.*, "Deep android malware detection," in *Proceedings of the seventh ACM on conference on data and application security and privacy*, 2017, pp. 301–308.

[37] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-an, and H. Ye, "Significant permission identification for machine-learning-based android malware detection," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 7, pp. 3216–3225, 2018.

[38] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon, "Androzoo: Collecting millions of android apps for the research community," in *Proceedings of the 13th international conference on mining software repositories*, 2016, pp. 468–471.

[39] D. Li, Z. Wang, and Y. Xue, "Fine-grained android malware detection based on deep learning," in *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2018, pp. 1–2.

[40] R. Vinayakumar, K. Soman, P. Poornachandran, and S. Sachin Kumar, "Detecting android malware using long short-term memory (lstm)," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 3, pp. 1277–1288, 2018.

[41] X. Su, D. Zhang, W. Li, and K. Zhao, "A deep learning approach to android malware feature learning and detection," in *2016 IEEE Trustcom/BigDataSE/ISPA*. IEEE, 2016, pp. 244–251.

[42] S. Hou, A. Saas, L. Chen, and Y. Ye, "Deep4maldroid: A deep learning framework for android malware detection based on linux kernel system call graphs," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW)*. IEEE, 2016, pp. 104–111.

[43] Z. Chen, J. Liu, Y. Shen, M. Simsek, B. Kantarci, H. T. Mouftah, and P. Djukic, "Machine learning-enabled iot security: Open issues and challenges under advanced persistent threats," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[44] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "Ddos in the iot: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017.

[45] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-baiot—network-based detection of iot botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018.

[46] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[47] A. R. Gad, A. A. Nashat, and T. M. Barkat, "Intrusion detection system using machine learning for vehicular ad hoc networks based on ton-iot dataset," *IEEE Access*, vol. 9, pp. 142 206–142 217, 2021.

[48] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40 281–40 306, 2022.

[49] S. Aurangzeb, M. Aleem, M. A. Iqbal, M. A. Islam *et al.*, "Ransomware: a survey and trends," *Journal of Information Assurance & Security*, vol. 6, no. 2, pp. 48–58, 2017.