
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 3, pp. 531–549, DOI 10.12694/scpe.v24i3.2155

BIG DATA IN HEALTHCARE – A COMPREHENSIVE BIBLIOMETRIC ANALYSIS OF CURRENT RESEARCH TRENDS

AIJAZ AHMAD RESHI*, ARIF SHAH†, SHABANA SHAFI‡ AND MAJID HUSSAIN QADRI§

Abstract. The primary purpose of this study is to perform a comprehensive bibliometric analysis of research landscape of big data in healthcare. Big data as a significant technology used in healthcare during the past decade has led to the exponential growth in scientific literature. This study is focused on analysis of many crucial bibliometric indicators such as, overall research output, author productivity, institutional productivity, country wise productivity, collaboration analysis, research trends along with a thematic focus of research output in big data and healthcare. The analysis has been performed on 2294 research articles published in 1018 publication sources from SCOPUS and Web of Science databases. The initial results of the study performed from year 2012 reveals that in the first year 6 research articles were published in the given domain. Then every year the growth of published articles in the field was exponential, however years 2019 to 2021 were the most productive and incremental in terms of number of publications. The analysis results of the study present the performance analysis of research production in terms of annual scientific production, most globally cited articles, author's production over the time, most productive countries, and most relevant affiliations. In addition, the science mapping analysis including the indicators such as, keyword Co-occurrence, Thematic Mapping, Most Relevant Authors, annual source distribution, and collaboration Network analysis has been presented. The study delivers expedient contribution to the field of study by noticeably offering comprehensive analysis results regarding research hotspots and trends, thematic emphasis, and future direction of research in the field. These outcomes will aid researchers in big data and healthcare in planning and designing the research and the challenges and opportunities needed to be explored.

Key words: bibliometric analysis, big data, healthcare, thematic mapping, KCN, Collaboration networks

1. Introduction. Knowledge not only aids in better understanding a subject or a situation, but it also helps in making better decisions. The more data leads to the better decisions and the better results. In the present technological era where almost everything is digitized [1], a large number of applications create a huge amount of data. International Data Corporation (IDC) estimates that the global data sphere will reach 175 Zettabytes by 2025, in its Data Age 2025 research for Seagate[2]. This situation pushed technocrats and large companies to design the solutions which not only manage the data but also retrieve meaningful information out of that data. Eventually these solutions should fulfill both current and future demands irrespective of the domains. There is no doubt that healthcare organizations have major contribution in data generation, making the industry extremely data-intensive, and electronic data has played an increasingly important role in understanding and improving health [3]. Like other application domains, healthcare too has varieties of data which makes big data in healthcare intimidating and the velocity of healthcare data demands effective and efficient data management solutions. Big data scientists are in high demand because they can assist in uncovering relationships and analyzing patterns and trends in data. Big data analytics has the potential to enhance care, save lives, and reduce costs. As a result, big data analytics apps in healthcare take advantage of the massive amount of data available to extract insights and make better decisions [4]. Big data will assist the healthcare industry in developing more comprehensive and insightful diagnosis and treatments, resulting in greater quality care at lower costs and better overall outcomes.

Thus the enormous need for big data has prompted researchers, academics, and professionals to delve deeper into it, as seen by the large number of research studies performed in the near past at a quick pace. It

*Department of Computer Science, College of Computer Science and Engineering, Taibah University, Al Madinah, al Munawarah, Saudi Arabia (aijazonnet@gmail.com).
†Deparment of Management, OPJS University, Rajasthan, India.
‡Department of Computer Science, College of Computer Science and Engineering, Taibah University, Al Madinah, al Munawarah, Saudi Arabia.
§Department of Management Studies, University of Kashmir.

is getting increasingly difficult to keep up with the updated body of knowledge in the field of study as the research outputs are published at an exponential rate [5]. The objective of this study is to combine existing references with bibliometric analysis to show the status and development of research in big data in healthcare, as bibliometrics is one of the best approaches that use statistical tools to assess research articles and other types of publications over time.

To achieve the primary goals of this study performance analysis and scientific mapping analysis have been performed in this study. Many important bibliometric indicators have been analyzed such as, scientific article production, total citations per year, most productive authors, most internationally cited documents, most cited nations, most relevant affiliations, most relevant keywords, and most relevant authors are all included in the performance analysis. In addition science mapping analysis has been carried out to analyze the bibliometric indicators such as, keyword co-occurrence, coupling map, theme mapping, and author, institution and country collaboration networks.

Highlights:

- To provide a comprehensive bibliometric analysis reference workflow for an interdisciplinary field of study.
- To provide the analysis of evolution and research trends of a field of study such as big data and healthcare.
- To analyze the research productivity in terms of countries, Institutions and authors in the field.
- To analyze the collaborative relationships among countries, institutions and authors in the research field.
- To provide science mapping and thematic analysis for big data and healthcare domain.

2. Related Work. In the present technological era every hour, a massive amount of data is generated from almost every environment, big or small. The data is considered a crucial asset of every organization and field. In other words, the more detailed data we have, the more optimally the operations and decisions can be made by the organization, Aria et al. [5]. Considering such critical importance of data, we are flooded by massive amounts of data from every organization. Information and communication technology (ICT) advances have significantly contributed towards generating and collecting the amounts of data to an extreme where it became very challenging to manage it. To deal with the massive amounts of data related to research studies, their results, contributions, and impact, technologies such as big data analysis and Text mining have been effectively used in recent years. Text mining is a method of analyzing a collection of documents as a knowledge-intensive task to recognize and discover interesting patterns Guandong Son et al. [6]. Bibliometrics is a quantitative information assessment method to analyze the emerging developments in a desired field of study to find measurable research output.

Guo, Yuqi, et al. in [7] have proposed a bibliometric search strategy to provide critical insights in the research related to the application of Artificial Intelligence in health care. The study has used the databases of Web of Science, such as the Science Citation Index (SCI) and the Social Science Citation (SSCI). The authors have conducted the temporal and spatial bibliometric analysis. In addition, the analysis has been done based on word co-occurrence, co-country, and co-authorship.

Dash, S et al. [8] provide a comprehensive review of big data management and analysis in healthcare. The results of the research review have proven the potential of big data application in better clinical prognosis. In addition, the application of big data analysis has been proven as an effective tool in reducing cost in analytics, making clinical decision support systems, making superior strategies of treatment as well as detecting and avoiding the fraudulent associated with data.

Galetsi, P et al. [9] present the applicability of big data Analytics in healthcare in terms of nature and scale of innovations in information processing and analysis tools. The study also explored the impact of technological tools and the possible information sources. To achieve the goals, bibliometric analysis has been done on the data sources extracted from Web of Science and Scopus databases. The findings of the study report a massive amount of work done published in numerous research papers related to oncology and neurology. The study thus reports the possible usability of big data analytics in advanced health information and Decision support systems. The findings further prove that the analysis tools will provide the solutions for sophisticated disease prognosis and diagnosis systems.

The application of data mining in healthcare and more specifically in medicine has been extensively researched in recent times throughout the world Hu, Yuanzhang et al. [10]. There is a significant need for research to provide a clear picture of growth and development trends of data mining and analytics in the field using bibliometrics. The study has proposed a bibliometric analysis example to give a clear overview of applying data mining methods in medicine. Various visualization tools have been used to analyze the citations, research collaborations, and spatial dissemination related to data mining in medicine.

Ale Ebrahim S et al. [11] have used a bibliographic method to study the research trend in drug delivery. The data has been sourced from SCOPUS to examine the research trend for almost 45 years till 2019. The network analysis method has been used by the study for research output analysis. The bibliographic analysis has been done on journal research articles in terms of citations, country-wise contribution based on specific keywords and topics. The bibliometric reports have been used to study the present status and required upcoming research directions in drug delivery.

Wu Haiyang et al. [12] have performed a study aimed at the identification of research output and impact of ultrasound micro bubble as a therapeutic method for diseases such as cancer and neurological and cardiovascular disorders. The literature related to the field of study has been used for 20 years till 2019 extracted from sources like Web of Science Core collection and also SCOPUS data has been used for invalidation. The bibliometric results such as a number of publications, document citations, journal citations, H-Index, authorship, co-authorship, country wise, institution wise, and keyword-based analysis have been performed using various data processing and visualization tools. The study has reported an increased amount of research results supporting the applicability of the ultrasound bubble in worldwide trends generally, however, the United States is leading in the field. The trends in the study depicted the significant research needs for the possibility of ultrasound usability in drug delivery along with its diagnostic application.

A Similar study on the usage of Nano magnetic iron oxides in drug delivery for cancer therapies has been performed by Darroudi, M. et al. [13]. The bibliographic analysis has been performed on research trends in the field. The analysis results of the study depict the progressive usage of Nano magnetic iron oxides in drug delivery. The analysis has reported the potential applications and challenges involved in the usage of these carriers along with future research directions in general multidisciplinary domains and more specifically in the treatment of colorectal cancer.

Raban et al. [14] have performed another bibliometric analysis to study the evolutionary trends of two revolutionary data analytics fields, big data, and data sciences. The study reports a significant number of publications in big data along with a continuously increasing growth in data science research publications. In addition, the results have proven a recent emergence of publications with combined applications of both big data and data science. The study evaluates the bibliometric indicators such as fields of study, journal indicators, country of origin and funding, citation indices, and authorship.

Borges do Nascimento IJ et al. [15] have performed a study for the assessment of significance for the application of big data analytics in healthcare. The study is focused on core health indicators and primacies according to the World Health Organization's General program and European program of Work. The study attempted to identify the challenges and potential opportunities of analytical tools in relation to public health. To search the data for systematic review, six databases such as Web of Science, MEDLINE, SCOPUS Embase, Cochrane Database of Systematic Reviews, and Epistemonikos have been included in the study. The core objective of the systematic reviews of the study is to assess the impact of big data analytics on people's health indicators. The study has reported that big data analytics have provided diagnosis or prediction of diseases like diabetes and mental disorders along with diagnosis or prediction of some chronic diseases with accuracy ranging from moderate to very high.

3. Methodology. Bibliometric methods including science mapping and performance analysis provide valuable insights about the evolution of a specific Field. [16]. Systematic Literature Review (SLR) has been used in this study to retrieve literature for the bibliometric analysis. SLR is carried out using the set of procedures for searching numerous databases using a predefined search strategy, which improves the transparency, scientific, and comprehensiveness of the study [18]. A thorough systematic literature evaluation not only improves the dependability of the research, but it also eliminates the possibility of including studies that are irrelevant [19]. The focus of this study is to explore the research output published by research studies done under the theme

of 'big data and healthcare'. The study has used Scopus and Web of Science (WOS) databases to search the existing relevant literature on the topics related to the big data in healthcare. There are various reasons behind choosing the Scopus and WOS databases for retrieving the literature during this study.

1. Scopus and WOS are the most comprehensive databases available [20] [21].
2. Both databases are updated on a regular basis [22].
3. Both databases are adaptive in terms of debugging and data processing [23].

The SLR has been conducted using a precise but rigorous methodology. The search parameters and search refinement plays a key role in bibliometric analysis, thus needing a significant focus. Keeping this in mind, the process has been broken into three primary stages, commencing with database search in the chosen databases. In the second stage, search criteria were applied to determine which studies should be included for further analysis. Finally, bibliometric analysis has been performed to examine the scientific production of articles, and the results have been presented for performance analysis and science mapping analysis. figure 3.1 presents the prism flow diagram of the approach adopted in the study.

3.1. Literature search. In the literature search phase the same filtration criteria have been used for both the databases. The primary keywords such as, "big data" and "healthcare" have been used to improve the search criteria linked with big data and healthcare. While setting a search criteria the conditional params like AND, OR are used when two or more keywords are used in the query. In this study AND helper has been used in the middle of two keywords. As a result, the search terms became "big data" AND "healthcare", implying that both terms shall be regarded equally while searching. Thus the search criteria resulted in a list that included articles and information related to big data and healthcare. The search criteria had been further narrowed down with numerous parameters such as document type, publishing stage, and language to receive more polished and relevant data.

3.2. Study selection. Since the literature search in this study has been restricted to only publications present in Scopus and Web of Science databases. To ensure that the data is enough and relevant, three filtration parameters such as Document type as 'articles', stage as 'published' and Language as 'English' have been used. The main purpose of applying the language filter is to keep the focus only on a single language. The literature search resulted in 4189 articles from Scopus and 1676 articles from Web of Science during the initial keyword search. In the next step 470 articles were eliminated from Web of Science and 2262 articles from Scopus by applying the filter's parameters, leaving 2732 articles for further analysis. The information has been exported in BibTeX format from Scopus, while the data has been downloaded in plain text format from Web of Science. To ensure that the dataset records are unique, the duplicate articles have been searched, which resulted in a total of 839 duplicate articles. These duplicate documents have been dropped from the final dataset; the remaining 2294 articles have been used in further analysis of the study. The detailed description about the data used in the study has been shown in Table 3.1. Finally, the Biblioshiny a web based tool for Bibliometrix, an R based bibliometric analysis platform has been used to analyze the most significant bibliometric parameters. These parameters include the author contributions, journals, countries and institutions, exploring research hot spots, research trends and forecasting the in the selected research domain.

4. Performance Analysis.

4.1. Annual Scientific Production. Despite the fact that big data has been there since the 1990s, it became popular when data began to rise at a rapid rate. Big data in healthcare has attracted a large number of researchers from all over the world, with an annual growth rate of 18.59 percent. The path of publication on big data in and healthcare began in 2012, as seen in Figure 4.1. Table 4.1 shows that in 2012, six research articles were published, with an average total citation of 8.33 for each publication. In 2013, the number of studies increased to 15, and the total number of citations per article increased dramatically to 83.46, with a mean citation of 9.27. Since then, we've seen a significant increase with each passing year. For example, in 2014, there were 57 publications, which was three times the prior year, and we saw the same pattern in the following years, with 110 publications in 2015 and 172 publications in 2016. We did see, however, that the average total citation per article fluctuated a little and that the average total citation per year fell. In 2017, there was relatively little growth compared to 2016, with 181 publications and an average total citation per piece of 32.18. For the fourth year in a row, researchers have focused their emphasis on big in healthcare, resulting in 332, 447,

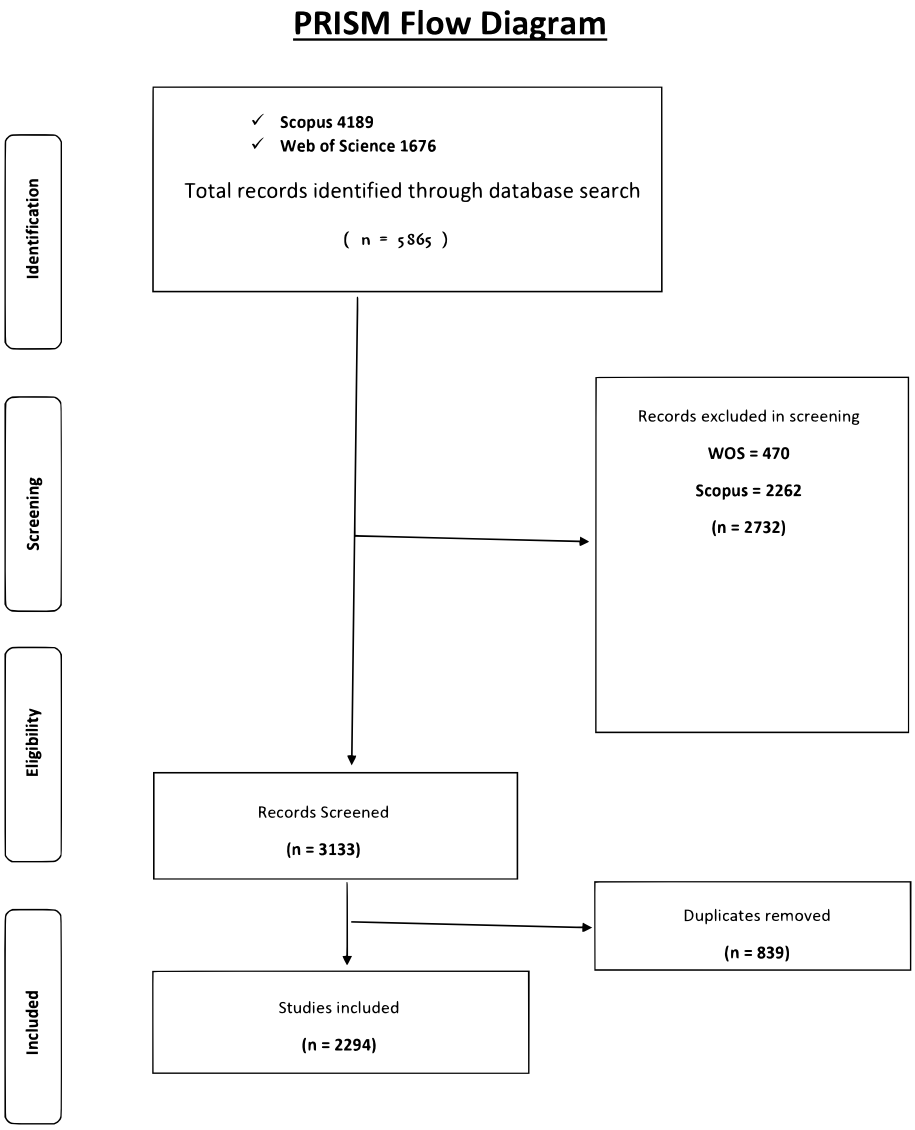

Fig. 3.1: Prism Flow Diagram of the study

468, and 417 papers. It's worth noting that 33 studies were published in January, indicating that 2022 will provide many publications.

4.2. Most Globally cited documents. According to bibliometric analysis of data from Scopus and Web Of Science (WOS), Rajkomar's work titled "Scalable and accurate deep learning using electronic health records" is the most cited, with 614 total citations and 122.8 total citations per year. Predictive modeling of electronic health record (EHR) data is expected to drive Customized therapy and enhance healthcare quality, according to this article published in 2018. With 433 total citations and 72.167 total citations per year, LEE JG's study on Deep Learning in Medical Imaging remains a highly referenced publication. In 2017, LEE JG presented a study that emphasized the relevance of artificial neural networks (ANN) and the use of big data in ANN.

Table 4.2 demonstrates that articles published in 2016, 2017, and 2018 are highly referenced, indicating that the study is of high quality. Farahani B et al. wrote a study in 2018 called 'Towards fog-driven IoT eHealth: Promises and Challenges of IoT in Medicine and Healthcare,' which underlines the relevance of the

Table 3.1: Data Description

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2045:42:00
Sources (Journals, Books, etc)	1018
Documents	2294
Average years from publication	3.3
Average citations per documents	15.93
Average citations per year per doc	3.387
References	10600
DOCUMENT CONTENTS	
Keywords Plus (ID)	0
Author's Keywords (DE)	5465
AUTHORS	
Authors	7782
Author Appearances	10344
Authors of single-authored documents	196
Authors of multi-authored documents	7586
AUTHORS COLLABORATION	
Single-authored documents	209
Documents per Author	0.295
Authors per Document	3.39
Co-Authors per Documents	4.51
Collaboration Index	3.64

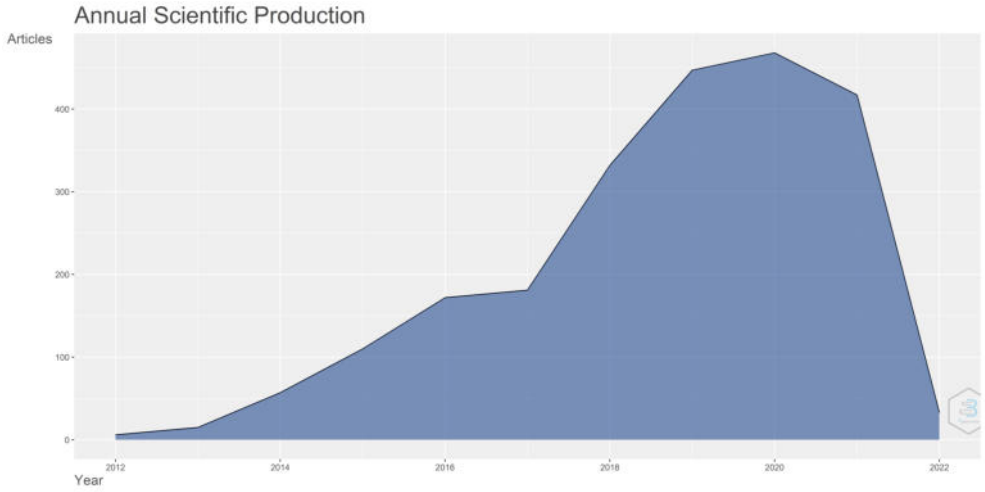

Fig. 4.1: Annual Scientific Productions

Internet of Things (IoT) and how it may help us move from clinic-centric to patient-centric healthcare. Big Data, according to Farahani B et al, is the key to revolutionizing the healthcare environment. WANG YC's work 'Big data analytics: Understanding its capabilities and potential advantages for healthcare organizations from 2018 is one of the most referenced, with 376 total citations, followed by Dimitrov's paper from 2016 with 370 total citations. Papers by Mohr DC, Vaishya R, Tu Yf, Yin s, Yin S, Zhnag Y, and Farahani B also received 300 or more citations, with Vaishya R's work receiving the most total citations per year with 115.667. Other works have received between 180 and 300 total citations.

Table 4.1: Annual total citations per year

S.no.	Year	Publications	MTCperArt*	MTCperYear*	Citable Years
1	2012	6	8.333333	0.833333	10
2	2013	15	83.46667	9.274074	9
3	2014	57	26.68421	3.335526	8
4	2015	110	31.30909	4.472727	7
5	2016	172	26.47674	4.412791	6
6	2017	181	32.18232	6.436464	5
7	2018	332	24.12651	6.031627	4
8	2019	447	13.12081	4.373602	3
9	2020	468	10.50641	5.253205	2
10	2021	417	2.455635	2.455635	1
11	2022	33	0.333333		0

*MTCperYear = Mean total citations per year MTCperArt = Mean total citations per article

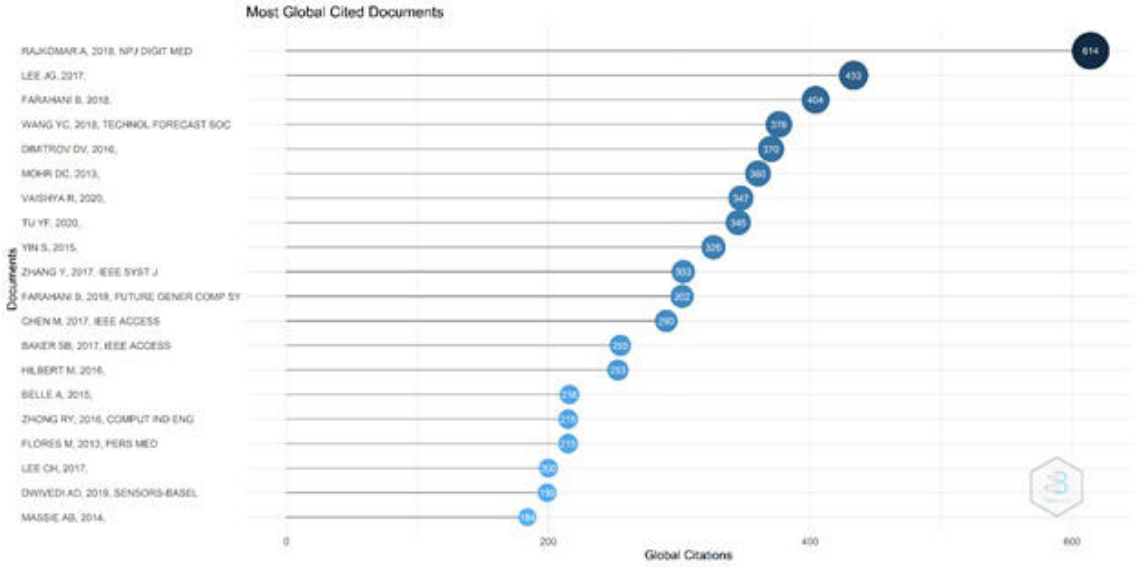

Fig. 4.2: Most globally cited documents

4.3. Author's Production over the Time. According to a bibliometric examination of data collected from 2012 to 2022, Zhang Y is the most productive author throughout that time span. He has 614 total citations and averages 120.8 total citations per year. Zhang Yi has over 80 articles and an h-index of 114. With around 433 total citations and 72 total citations per year, Lee H is the second most productive author of big data in healthcare, according to Fig 4.2 On the list of most productive authors, Wang Yichuan is ranked third. One of the greatest of his many writings is 'Understanding its capabilities and possible advantages for healthcare organizations.' Wang Yichuan discussed the potential benefits of big data in healthcare and how big data analytics may be used as a major pillar of the healthcare ecosystem in this study. Wang Yichuan has around 376 total citations, with an average of more than 75 each year. Chen M, Chen J, LI Y, Hossain M are few top authors who have contributed to the big healthcare theme.

We can refer to Fig 4.3 which highlights authors' production over the time. In this figure each article is denoted by a circular node. The size of the circular node denotes the number of the articles; the total citation is denoted by color. Darker the color means more citations. As per the figure Kim J, Lee S, Chen M and Zhang J are among few researchers who had contributed to this theme from the early period of 2014.

Table 4.2: Authors production

S. no	Author	Paper	Total Citations	TC per Year
1	RAJKOMAR A, 2018, NPJ DIGIT MED	Scalable and accurate deep learning with electronic health records	614	122.8
2	LEE JG, 2017,	Deep Learning in Medical Imaging: General Overview	433	72.167
3	FARAHANI B, 2018,	Towards fog-driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare	404	80.8
4	WANG YC, 2018, TECHNOL FORECAST SOC	Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations	376	75.2
5	DIMITROV DV, 2016,	Medical Internet of Things and Big Data in Healthcare	370	52.857
6	MOHR DC, 2013,	Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health	360	36
7	VAISHYA R, 2020,	Artificial Intelligence (AI) applications for COVID-19 pandemic	347	115.667
8	TU YF, 2020,	A Review of SARS-CoV-2 and the Ongoing Clinical Trials	345	115
9	YIN S, 2015,	Big Data for Modern Industry: Challenges and Trends	326	40.75
10	ZHANG Y, 2017, IEEE SYST J	Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data	303	50.5
11	FARAHANI B, 2018, FUTURE GENER COMP SY	Towards fog driven IoT eHealth: Promises and challenges of IoT in medicine and healthcare	302	60.4
12	CHEN M, 2017, IEEE ACCESS	Disease Prediction by Machine Learning Over Big Data from Healthcare Communities	290	48.333
13	BAKER SB, 2017, IEEE ACCESS	Internet of Things for Smart Healthcare: Technologies, Challenges, and Opportunities	255	42.5
14	HILBERT M, 2016,	Big Data for Development: A Review of Promises and Challenges	253	36.143
15	BELLE A, 2015,	Big Data Analytics in Healthcare	216	27
16	ZHONG RY, 2016, COMPUT IND ENG	Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives	215	30.714
17	FLORES M, 2013, PERS MED	P4 medicine: how systems medicine will transform the healthcare sector and society	215	21.5
18	LEE CH, 2017,	Medical big data: promise and challenges	200	33.333
19	DWIVEDI AD, 2019, SENSORS-BASEL	A Decentralized Privacy-Preserving Healthcare Block chain for IoT	199	49.75
20	MASSIE AB, 2014,	Big Data in Organ Transplantation: Registries and Administrative Claims	184	20.444

4.4. The Most Productive Countries. The distribution of publications and citations among the nations that contributed the most to the field from 2012 to 2012 will now be explored. According to the bibliometric study, the United States has contributed the most articles to the big data in healthcare issue, with 1223, which is 100 times or more than the other nations on the list. Table 4.3 reveals that the United States has 8987 total citations and 21.552 average article citations, placing it first in both total and average article citations. China is the second country to contribute to this subject, with 649 articles with 4791 citations and an average of 18.788 citations per article. With 509 publications and 2804 total citations, India is the third most productive country in the big data in healthcare field, raising issues about the quality of the research done by Indian experts. The United Kingdom, North Korea, Spain, Italy, France, Australia, and Canada are just a handful of the many countries that have contributed to the hot issue of big data in healthcare. Fig 4.4 shows that large healthcare is something that has prompted practically every government to investigate and do study. Every country, regardless of its economy or way of living, is interested in using big data into healthcare.

The color shade in Fig 4.4 is used to help us understand each nation's contribution to this revolutionary topic. For displaying the number of publications published by a country, light blue is used as the starting shade and dark blue is used as the higher end of the scale. We can clearly distinguish the United States of America, China, and India among the darker blue colored countries, indicating that they have donated the most, while Costa Rica, Iceland, Kazakhstan, and a few other countries have contributed relatively less.

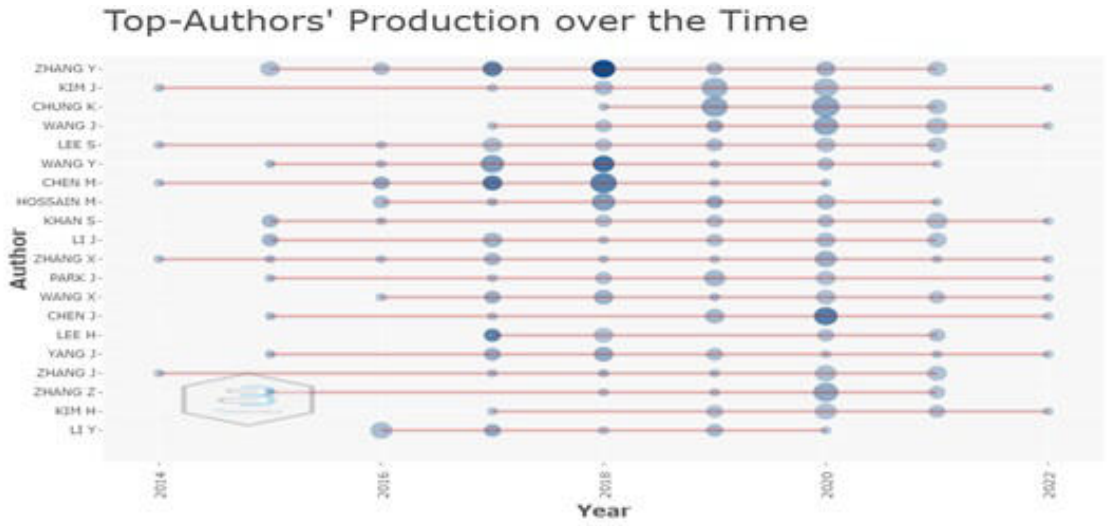

Fig. 4.3: Top-Author's Production over time

Table 4.3: Country-wise publications and citation rank

Rank	Country	Publications	Total Citations	Average Article Citation
1	USA	1223	8987	21.552
2	CHINA	649	4791	18.788
3	INDIA	509	2804	8.987
4	UK	377	1881	15.048
5	SOUTH KOREA	244	2117	16.802
6	SPAIN	186	444	8.377
7	ITALY	165	1265	17.095
8	FRANCE	151	730	16.977
9	AUSTRALIA	141	1511	18.887
10	CANADA	141	1034	19.148
11	SAUDI ARABIA	138	864	21.073
12	GERMANY	135	566	13.163
13	NETHERLANDS	127	419	12.697
14	JAPAN	112	167	6.185
15	PAKISTAN	99	262	13.1
16	BRAZIL	52	333	12.808
17	EGYPT	52	407	16.28
18	MALAYSIA	50	691	19.194
19	PORTUGAL	47	147	8.647
20	SINGAPORE	43	311	12.958

4.5. Most Relevant Affiliations. One of the most important components of bibliometric analysis is determining which institutions are the most prolific. Fig 4.5 shows that King Saud University is the most productive university in terms of contributions to big data in healthcare issues, with 55 papers. With 40 publications, Sanford University came in second. With 30 publications published, Harvard Medical School and Huazhong University of Science and Technology are ranked third on the list. With 28 publications published, Kyoncgi University rounds out the top five. University of Florida, Oxford University, University of Minnesota, National University of Singapore, and Sejong University are among the top 10 universities that have studied big data in the healthcare field.

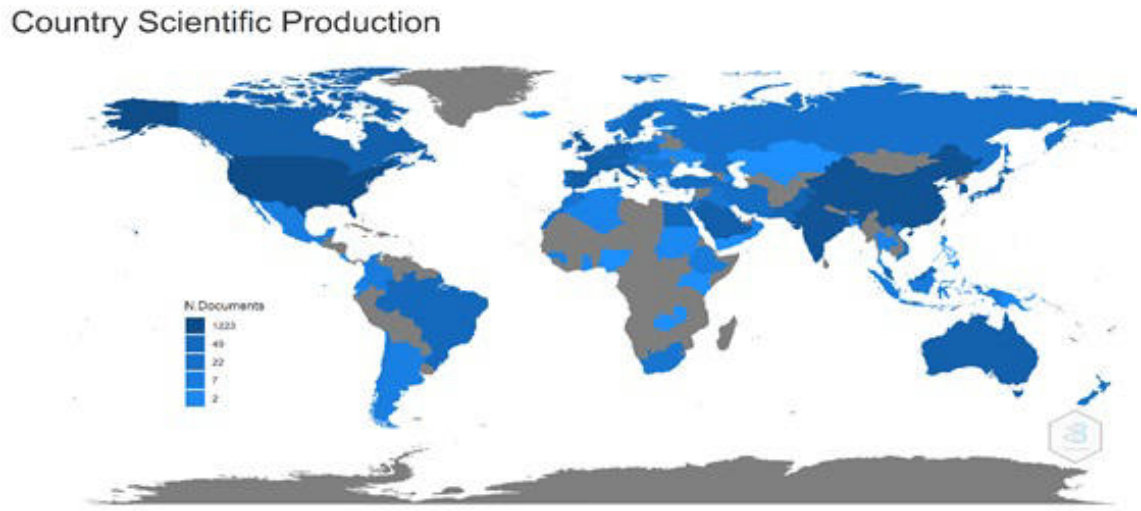

Fig. 4.4: Country-wise Scientific Production

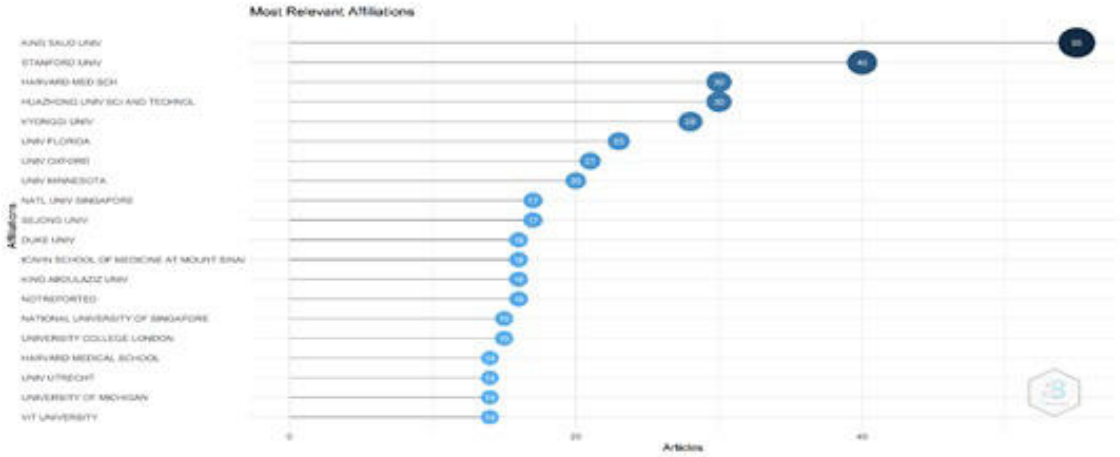

Fig. 4.5: Most relevant affiliations

5. Science mapping analysis.

5.1. Keyword Co-occurrence. Key word co-occurrence analysis is widely used in knowledge mapping. Systematic literature review for knowledge mapping of current state and future growth of scientific research of any study area is very crucial, but due to its manual nature it becomes very challenging and time consuming [24]. Fig 4.3 shows the number of occurrences of the keywords in the publications. There are 793 occurrences of the word 'big data', followed by 199 occurrences of the word 'healthcare'. The occurrences of other related words like 'machine learning', 'artificial intelligence', 'cloud computing', 'internet of things' and so on are also given in the graph in decreasing order of the occurrences. These results reveal that there is a significant rise in usage and togetherness between the two words, "big data" and "healthcare" followed by the words representing the technologies very close to big data and healthcare. The word cloud of the primary and other related keywords is given in Fig 5.1 and the number of most relevant word occurrences has been shown in Fig 5.2.

. To further perform the keyword co-occurrence analysis of the key words such as 'big data', healthcare", and other related keywords, a keyword co-occurrence network (KCN) has been created. KCN is represented

Fig. 5.1: Word Cloud

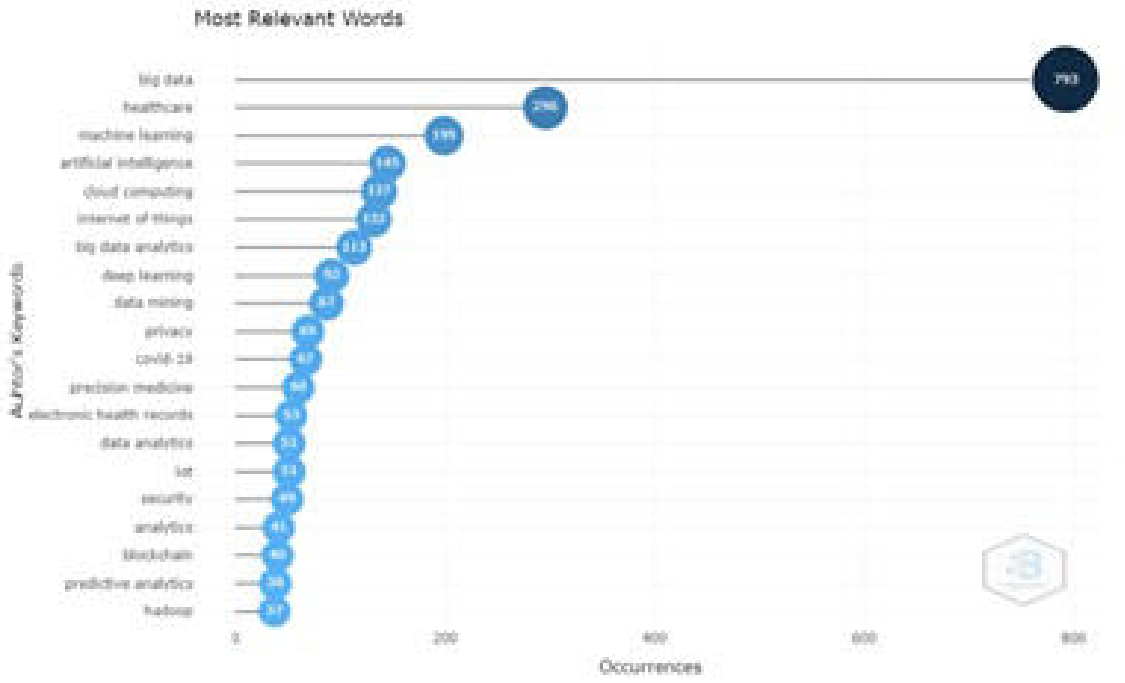

Fig. 5.2: Most Relevant Word occurances

as a graph with a set of nodes and edges. The keywords are represented as nodes and each co-occurrence is depicted by edges between the nodes. The weight of each edge connecting a pair of nodes is represented as the number of co-occurrences between the words representing the node pair [25]. The weights of the links in a KCN are visually shown as the thickness of each edge. The thickness of each edge thus is proportional to the corresponding weight of the edge. Fig 5.3 shows the KCN with three clusters represented by different colors as Red (Cluster1), Green (Cluster2), and Blue (Cluster3). The node size in the KCN represents the proportionality of frequency of words, bigger the size of node more is the occurrence of the word. As can be seen in the KCN two words in Cluster one such as, 'big data' followed by 'healthcare' are more frequent. Further the edge weight between these two words is thicker than all the other edges meaning more co-occurrences between these two

Fig. 5.3: Keyword Co-occurrence network

words in comparison to all other word co-occurrences.

5.2. Thematic mapping. Thematic mapping is a conceptual structure in bibliometric. The thematic mapping outlines the conceptual structure of the keywords under consideration [26]. Thematic mapping visualizes the theme structure in the form of four quadrants of a thematic map, each quadrant representing a theme as shown in Fig 5.4. The themes are categorized in two properties such as density and centrality. Density represented by vertical axis is the degree of correlation of keywords while centrality represented by horizontal axis measures the cohesiveness among the keywords. The thematic map in the figure provides the analysis of big data and healthcare. The map is illustrating the themes into four quadrants, the upper right quadrant represents motor theme, upper left quadrant represents niche theme, lower right corner represents the basic theme and lower left represents emerging or declining theme. Themes like big data, healthcare, cloud computing, AI, Machine Learning Deep learning are lying in the basic theme quadrant which are very vital in the development of the field of study. Since the thematic mapping has been done as Co-word analysis which identifies keyword clusters. They are regarded as themes, and their density and centrality being utilized to group themes and map them on a two-dimensional graph.

Themes like healthcare records, public health, data analytics and predictive analytics seen in niche quadrants depicted in the thematic map have established internal bonds. The thematic analysis thus reveals that the themes in niche quadrant such as data analytics and predictive analysis are imminent areas to be associated with big data and healthcare. Researchers in this field of study need to apply these methods to progress further in this field of study. The themes such as personal medicine, precision medicine and data science are sandwiched in all the four quadrants of the thematic map.

5.3. Most Relevant Authors. The results of the most relevant authors show the number of papers published by each author in the current research domain. The top 20 authors who published extensively have been shown in the results as a plot given in Fig 5.5. The numbers in the circles of the plot shows the number of publications published by each author. The results reveal that the highest number of publications, according to the Web of Science and SCOPUS database analysis; related to big data and healthcare have been contributed by ZHANG Y. The author is thus the most relevant author with 22 publications followed by KIM J with 20, which is also very close to the top author in terms of relevance of authors. CHUNG K with 19 publications in the domain is also very close. As can be seen in the plot, the least number of publications published in the

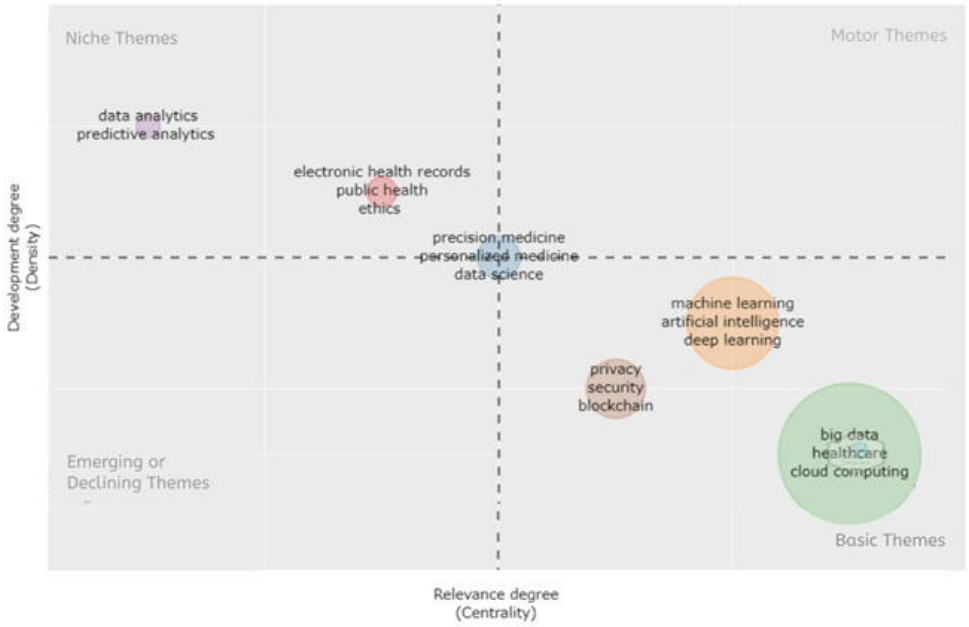

Fig. 5.4: Thematic Mapping

studied research domain are 10 in number, which is also a significant number in terms of relevance. The results thus reveal that a good number of researchers are currently working in the research area considered in this study. Further the results of Lotka's law analysis, given in Fig 5.5, which gives the frequency of publication by the relevant authors in big data and healthcare.

6. Annual Source Growth.

6.1. Annual Source Distribution. The annual source growth is another important bibliometric indicator analyzed in the study. This indicator depicts the growth distribution of publications related to the given domain by source. The results shown in Fig 6.1 depict the number of publications available in top 8 journals. The overall results reveal 2294 articles published by 1019 journal sources. According to the Bradford's Law analysis there are three zones: Zone1 includes 49 journals with 758 published articles (33.04%), Zone2 consisted of 252 journals, with 779 published articles (33.95%), and Zone3 consists 718 journals with 757 articles (32.99%). The Bradford analysis graph for top 50 sources is shown in Fig 6.3 the analysis data for the graph has been depicted in Table 6.1. The source distribution percentage has been illustrated in Fig 6.2.

6.2. Collaboration Network Analysis. Collaboration Networks are used to visualize the research collaborations between authors, research groups, institutions, and countries [27] [28]. Collaboration networks can thus be used to illustrate how intensely the researchers work together in a field of study.

6.3. Author Collaborations. The research output of studies is published as research articles; the co-author relationship of these articles determines the author collaborations. The author collaboration is a significant bibliometric indicator in studying the collaboration patterns of authors. The author's research collaboration analysis results of this study are shown in the form of Collaboration network in Fig 6.4. Each node in the collaboration network graph represents an author, the edges between the nodes depicts the collaboration relationship between the linked authors. The weight of each edge represented by edge thickness in the figure represents the number of articles co-authored by two researchers. The results in the figure thus reveal that Kim J. and Chung K. In the first cluster drawn in red has the highest papers co-authored in the field of study. Similarly other author collaborations can be easily understood in the figure consisting of the author nodes with significant collaboration relationships.

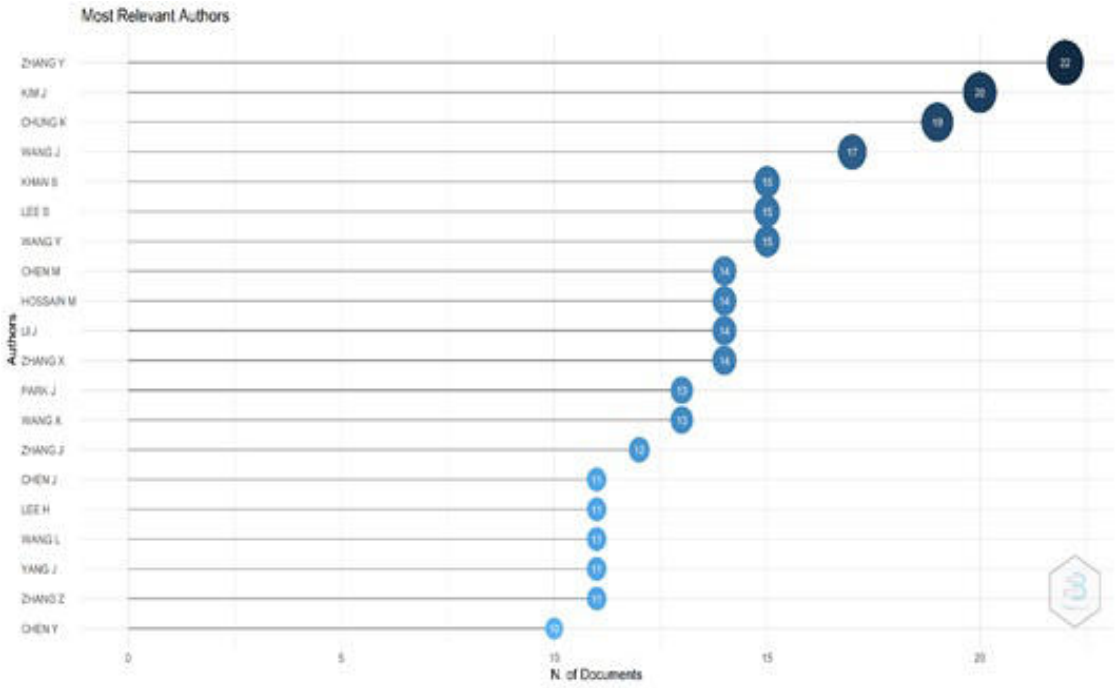

Fig. 5.5: Most Relevant Authors

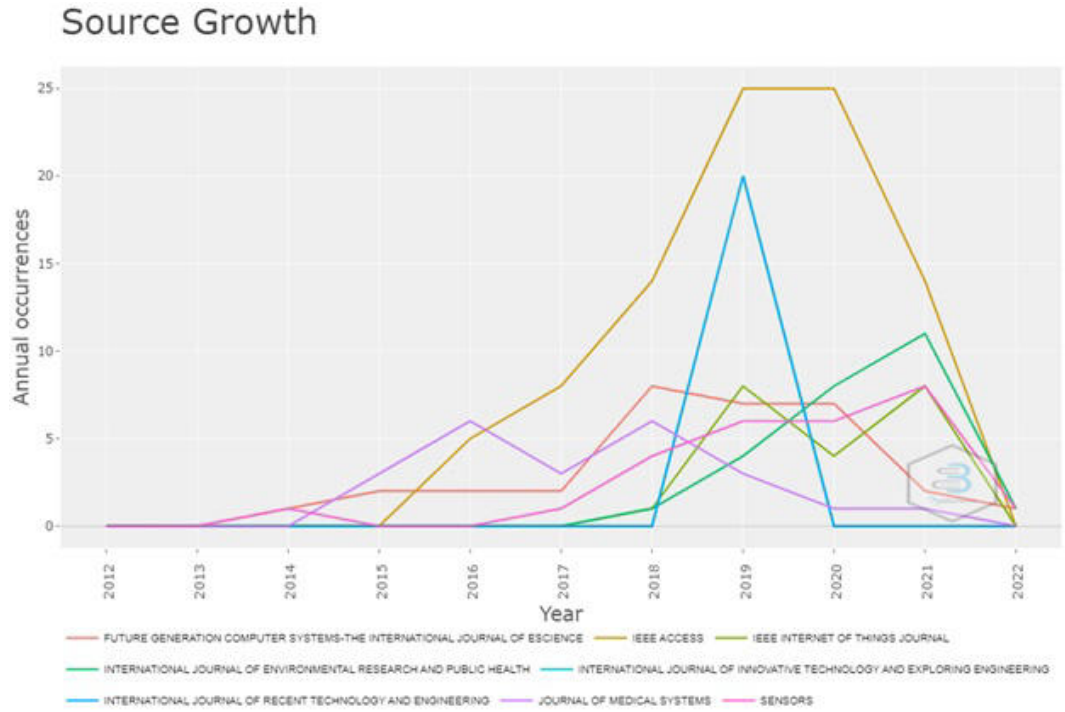

Fig. 6.1: Annual Source growth

Table 6.1: Source clustering through Bradford's Law

Rank	SO	Freq	cumFreq
1	IEEE ACCESS	91	91
2	FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE	32	123
3	SENSORS	27	150
4	INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH	25	175
5	JOURNAL OF MEDICAL SYSTEMS	23	198
6	IEEE INTERNET OF THINGS JOURNAL	21	219
7	INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND EXPLORING ENGINEERING	20	239
8	INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY AND ENGINEERING	20	259
9	JOURNAL OF ADVANCED RESEARCH IN DYNAMICAL AND CONTROL SYSTEMS	19	278
10	YEARBOOK OF MEDICAL INFORMATICS	19	297
11	BMC MEDICAL INFORMATICS AND DECISION MAKING	18	315
12	JOURNAL OF BIG DATA	18	333
13	INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS	17	350
14	TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE	17	367
15	JOURNAL OF AMBIENT INTELLIGENCE AND HUMANIZED COMPUTING	15	382
16	PLOS ONE	15	397
17	SUSTAINABILITY	15	412
18	BIG DATA	14	426
19	INTERNATIONAL JOURNAL OF ENGINEERING AND TECHNOLOGY(UAE)	14	440
20	JOURNAL OF COMPUTATIONAL AND THEORETICAL NANOSCIENCE	14	454
21	JOURNAL OF HEALTHCARE ENGINEERING	14	468
22	BMJ OPEN	13	481
23	JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION	13	494
24	APPLIED SCIENCES-BASEL	12	506
25	HEALTH AND TECHNOLOGY	12	518
26	HEALTHCARE INFORMATICS RESEARCH	12	530
27	IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS	12	542
28	INTERNATIONAL JOURNAL OF ENGINEERING AND ADVANCED TECHNOLOGY	12	554
29	INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT	12	566
30	JOURNAL OF SUPERCOMPUTING	12	578
31	BIG DATA RESEARCH	11	589
32	CLUSTER COMPUTING-THE JOURNAL OF NETWORKS SOFTWARE TOOLS AND APPLICATIONS	11	600
33	JOURNAL OF BIOMEDICAL INFORMATICS	11	611
34	MULTIMEDIA TOOLS AND APPLICATIONS	11	622
35	NPJ DIGITAL MEDICINE	11	633
36	WIRELESS PERSONAL COMMUNICATIONS	11	644
37	COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE	10	654
38	INTERNATIONAL JOURNAL OF SCIENTIFIC AND TECHNOLOGY RESEARCH	10	664
39	CIN-COMPUTERS INFORMATICS NURSING	9	673
40	FUTURE GENERATION COMPUTER SYSTEMS	9	682
41	IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS	9	691
42	INFORMATION SYSTEMS FRONTIERS	9	700
43	INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS	9	709
44	INTERNATIONAL JOURNAL OF APPLIED ENGINEERING RESEARCH	9	718
45	APPLIED SCIENCES (SWITZERLAND)	8	726
46	HEALTH INFORMATICS JOURNAL	8	734
47	INTERNATIONAL JOURNAL OF ADVANCED SCIENCE AND TECHNOLOGY	8	742
48	INTERNATIONAL JOURNAL OF HEALTHCARE MANAGEMENT	8	750
49	JOURNAL OF MEDICAL IMAGING AND HEALTH INFORMATICS	8	758
50	PERSONALIZED MEDICINE	8	766

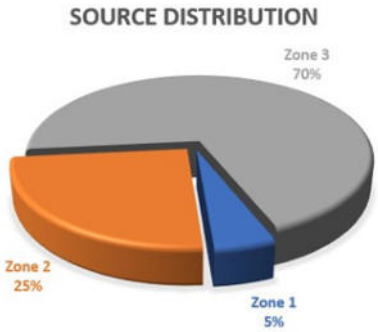

Fig. 6.2: Source distribution percentage

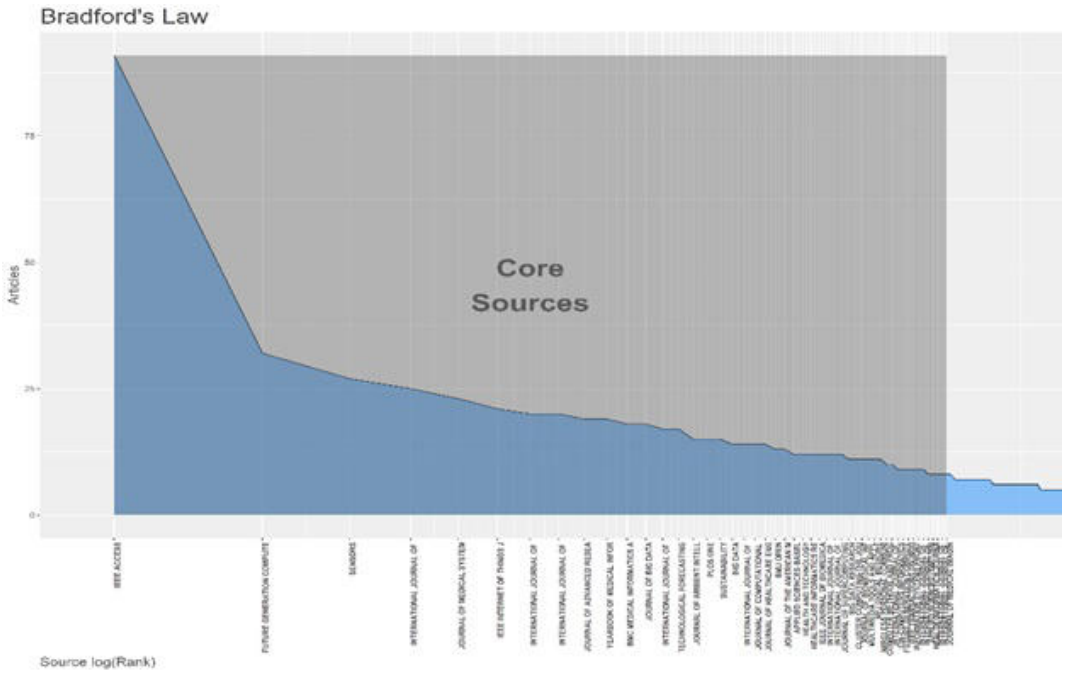

Fig. 6.3: Core Sources

6.4. Country Collaborations. In addition to author collaborations, the collaboration relationship between different countries who have significantly worked in big data and healthcare has also been analyzed in this study. The collaboration network representing the collaboration relationship has been presented as a graph in Fig 6.5. The results reveal that the most significant countries in terms of research collaboration in big data and healthcare are the USA (with betweenness = 177.83, closeness = 0.017) and China (with betweenness = 95.11, closeness = 0.014).

6.5. Institutional Collaborations. Institutional collaboration analysis is also considered as one of the important bibliometric indicators. The graph in Fig 6.6 visualizes the collaboration relationship between the institutions. The centrality measure in the collaboration network graph is indicated by the size of the node. The King Saud University is leading the centrality and betweenness, followed by Huazhong University Science and Technology, Taif University, and Itmo University. King Saud University has a closeness score of 0.0039

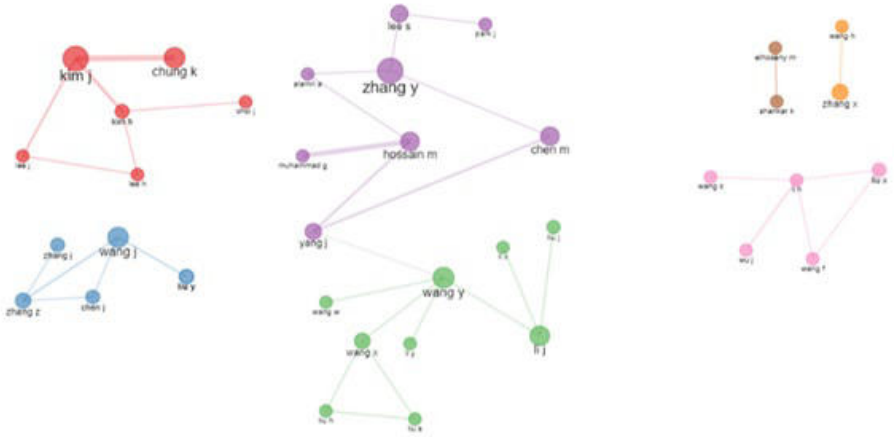

Fig. 6.4: Author Collaborations Network

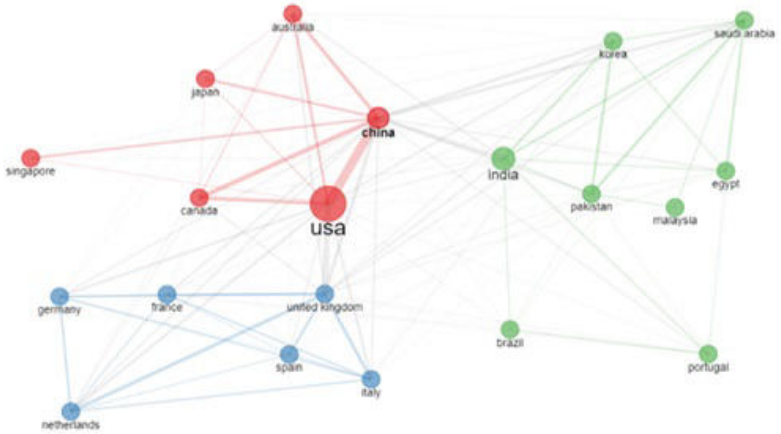

Fig. 6.5: Country Collaborations Network

while the other three universities share the equal closeness score of 0.0038.

7. Conclusion. This study is aimed to perform a comprehensive bibliometric analysis of the research landscape of big data in healthcare. Big data was effectively used in many healthcare application domains during the past decade which led to the exponential growth of the number of research studies in the field [29][30]. This study is focused on analysis of many crucial bibliometric indicators in an interdisciplinary research domain such as big data and healthcare. One of the contributions of this study is to identify the research studies and their reported outputs in the field of big data in healthcare. The study will be used as a reference by the research community, editors of research journals, industry professionals and academicians to recognize the recent state of scientific research, most prominent research articles, most prolific researchers, most prominent sources and potential collaborations between authors, institutions and countries in the field. The study reviews and summarizes the scientific literature, identifies the future research trends and directions.

A total of 5685 articles were extracted from the Scopus and web of science databases out of which 2294 were selected for further analysis in this study. The study contributes prominently to the body of research. The

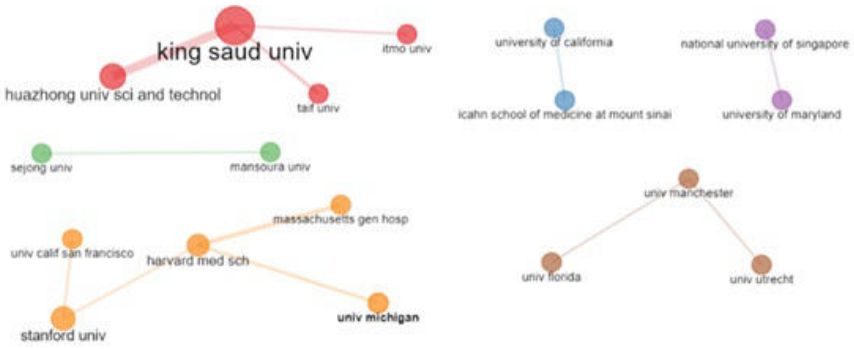

Fig. 6.6: Institutional Collaborations Network

results of the study reveal that the most relevant publishing source was IEEE ACCESS. Further, the analysis results reveal that ZHANG Y, KIM J. are the most relevant authors with the highest number of publications in the study domain. Similarly, the most relevant country which contributed to the research field over the last decade in terms of number of publications is the USA. In the institutional contribution analysis, King Saud University is the most productive university in terms of contributions to big data in healthcare followed by Sanford University. In the thematic mapping, the results reveal that the big data in healthcare is significantly emerging along with the themes like, data analytics, predictive analytics, electronic health records, public health and are closely connected to the field of big data in healthcare.

Acknowledgments. Authors would like to thank their friends, colleagues and their respective institutions for support in conducting in this study.

REFERENCES

[1] Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis and future prospects. J Big Data 6, 54 (2019). https://doi.org/10.1186/s40537-019-0217-0

[2] IDC , The Digitization of the World From Edge to Core https://www.idc.com/getdoc.jsp?containerId=prUS47560321, accessed on : 09 Feb 2022

[3] Aijaz Ahmad Reshi, Furqan Rustam, Arif Mehmood, Abdulaziz Alhossan, Ziyad Alrabiah, Ajaz Ahmad, Hessa Alsuwailem, Gyu Sang Choi, An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification, Complexity, vol. 2021, Article ID 6621607, 12 pages, 2021. https://doi.org/10.1155/2021/6621607

[4] Reshi AA, Ashraf I, Rustam F, Shahzad HF, Mehmood A, Choi GS. Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms. PeerJ Computer Science 7:e547 https://doi.org/10.7717/peerj-cs.547

[5] Aria and Cuccurullo, 2017), A bibliometric analysis of pandemic and epidemic studies in economics: future agenda for COVID-19 research, https://www.sciencedirect.com/science/article/pii/S2590291121000619#bib2

[6] Guandong Song, Jiying Wu, Sihui Wang, Text Mining in Management Research: A Bibliometric Analysis", Security and Communication Networks, vol. 2021, Article ID 2270276, 15 pages, 2021. https://doi.org/10.1155/2021/2270276

[7] Guo, Yuqi, et al. Artificial Intelligence in Health Care: Bibliometric Analysis." Journal of medical Internet research vol. 22,7 e18228. 29 Jul. 2020, doi:10.2196/18228

[8] Dash, S., Shakyawar, S.K., Sharma, M. et al. Big data in healthcare: management, analysis, and prospects. J Big Data 6, 54 (2019). https://doi.org/10.1186/s40537-019-0217-0

[9] Galetsi, P. and Katsaliaki, K. (2020), Big data analytics in health: an overview and bibliometric study of research activity. Health Info Libr J, 37: 5-25. https://doi.org/10.1111/hir.12286

[10] Hu, Yuanzhang MDa; Yu, Zeyun MDB; Cheng, Xiaoen MDa,∗; Luo, Yue MDa; Wen, Chuanbiao MDa,∗ A bibliometric analysis and visualization of medical data mining research, Medicine: May 29, 2020 - Volume 99 - Issue 22 - p e20338 DOI: 10.1097//MD.0000000000020338

[11] Ale Ebrahim S., Zamani Pedram M., Ale Ebrahim N. (2020) Current Status of Systemic Drug Delivery Research: A Bibliometric Study. In: Lai WF. (eds) Systemic Delivery Technologies in Anti-Aging Medicine: Methods and Applications. Healthy Ageing and Longevity, vol 13. Springer, Cham. https://doi.org/10.1007/978-3-030-54490-4-2

[12] Wu Haiyang, Tong Linjian, Wang Yulin, Yan Hua, Sun Zhiming, Bibliometric Analysis of Global Research Trends on Ultrasound Microbubble: A Quickly Developing Field", Frontiers in Pharmacology, 2021, https://www.frontiersin.org/article/10.3389/fphar.2021.646626

[13] Darroudi, M., Gholami, M., Rezayi, M. et al. An overview and bibliometric analysis on the colorectal cancer therapy by magnetic functionalized nanoparticles for the responsive and targeted drug delivery. J Nanobiotechnol 19, 399 (2021). https://doi.org/10.1186/s12951-021-01150-6
[14] Raban, D.R., Gordon, A. The evolution of data science and big data research: A bibliometric analysis. Scientometrics 122, 1563–1581 (2020). https://doi.org/10.1007/s11192-020-03371-2
[15] Borges do Nascimento IJ, Marcolino MS, Abdulazeem HM, Weerasekara I, Azzopardi-Muscat N, Gonçalves MA, Novillo-Ortiz D Impact of Big Data Analytics on People's Health: Overview of Systematic Reviews and Recommendations for Future StudiesJ Med Internet Res 2021;23(4): e27275
[16] Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F (2011) Science mapping software tools: review, analysis, and cooperative study among tools. J Am Soc Inform Sci Technol 62:1382–1402. https://doi.org/10.1002/asi.21525
[17] Pittway, L. (2008) Systematic literature reviews. In Thorpe, R., Holt, R. The SAGE dictionary of qualitative management research. London: SAGE Publications.
[18] Tang, 2019, A systematic literature review and analysis on mobile apps in m-commerce: Implications for future research,Electronic Commerce Research and Applications doi:10.1016/j.elerap.2019.100885
[19] Mugomeri E, Bekele BS, Mafaesa M, et al (2017) A 30-year bibliometric analysis of research coverage on HIV and AIDS in Lesotho. Health Res Policy Syst 15:21. https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-017-0183-y
[20] Zyoud SH, Fuchs-Hanusch D (2017) A bibliometric-based survey on AHP and TOPSIS techniques. Expert Syst Appl 78:158–181. https://doi.org/10.1016/j.eswa.2017.02.016
[21] Borrett SR, Sheble L, Moody J, Anway EC (2018) Bibliometric review of ecological network analysis: 2010–2016. Ecol Model 382:63–82. https://doi.org/10.1016/j.ecolmodel.2018.04.020
[22] Aznar-Sánchez JA, Velasco-Muñoz JF, Belmonte-Ureña LJ, Manzano-Agugliaro F (2019) Innovation and technology for sustainable mining activity: a worldwide research assessment. J Clean Prod 221:38– 54. https://doi.org/10.1016/j.jclepro.2019.02.243
[23] Radhakrishnan S, Erbis S, Isaacs JA, Kamarthi S (2017) Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. PLoS ONE 12(3): e0172778. https://doi.org/10.1371/journal.pone.0172778
[24] Lozano, S., Calzada-Infante, L., Adenso-Díaz, B. et al. Complex network analysis of keywords co-occurrence in the recent efficiency analysis literature. Scientometrics 120, 609–629 (2019). https://doi.org/10.1007/s11192-019-03132-w
[25] Della Corte, V.; Del Gaudio, G.; Sepe, F.; Sciarelli, F. Sustainable Tourism in the Open Innovation Realm: A Bibliometric Analysis. Sustainability 2019, 11, 6114. https://doi.org/10.3390/su11216114
[26] Wu, W., Xie, Y., Liu, X., Gu, Y., Zhang, Y., Tu, X., Tan, X. (2019). Analysis of Scientific Collaboration Networks among Authors, Institutions, and Countries Studying Adolescent Myopia Prevention and Control: A Review Article. Iranian journal of public health, 48(4), 621–631
[27] A. A. Reshi, A. Alsaeedi and S. Shafi, Development and Web Performance Evaluation of Internet of Things testbed, 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-6, doi: 10.1109/ICCISci.2019.8716436.
[28] Gubbi J, et al. Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener Comput Syst. 2013;29(7):1645–60, https://doi.org/10.1016%2Fj.future.2013.01.010
[29] Batko, K., Ślęzak, A. The use of Big Data Analytics in healthcare. J Big Data 9, 3 (2022). https://doi.org/10.1186/s40537-021-00553-4
[30] Aria, M.; Cuccurullo, C.; D'Aniello, L.; Misuraca, M.; Spano, M. Thematic Analysis as a New Culturomic Tool: The Social Media Coverage on COVID-19 Pandemic in Italy. Sustainability 2022, 14, 3643. https://doi.org/10.3390/su14063643
[31] Reshi AA, Rustam F, Aljedaani W, Shafi S, Alhossan A, Alrabiah Z, Ahmad A, Alsuwailem H, Almangour TA, Alshammari MA, Lee E, Ashraf I. COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset. Healthcare. 2022
[32] Swain, B. K., Khan, M. Z., Chowdhary, C. L., Alsaeedi, A. SRC: Superior Robustness of COVID-19 Detection from Noisy Cough Data Using GFCC, Computer Systems Science and Engineering 2023, 46(2), 2337-2349
[33] Javaid, A., Siddique, M.A., Reshi, A.A. et al. Coal mining accident causes classification using voting-based hybrid classifier (VHC). J Ambient Intell Human Comput (2022)
[34] F. Rustam et al., COVID-19 Future Forecasting Using Supervised Machine Learning Models, in IEEE Access, vol. 8, pp. 101489-101499, 2020
[35] Khan, Y. F., Kaushik, B., Chowdhary, C. L., Srivastava, G. (2022). Ensemble Model for Diagnostic Classification of Alzheimer's Disease Based on Brain Anatomical Magnetic Resonance Imaging. Diagnostics, 12(12), 3193
[36] F. Rustam et al., Sensor-Based Human Activity Recognition Using Deep Stacked Multilayered Perceptron Model, in IEEE Access, vol. 8, pp. 218898-218910, 2020
[37] Shabana Shafi , Aijaz Ahmad Reshi and A. Kumaravel, "Wireless Sensor Network based Early Warning and Alert System for Radioactive Radiation Leakage", Middle-East Journal of Scientific Research 19 (12): 1602-1608, 2014
[38] Padinjappurathu Gopalan, S., Chowdhary, C. L., Iwendi, C., Farid, M. A., Ramasamy, L. K. (2022). An efficient and privacy-preserving scheme for disease prediction in modern healthcare systems. Sensors, 22(15), 5574

Edited by: Chiranji Lal Chowdhary
Special Issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Mar 19, 2023
Accepted: May 28, 2023