International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 13, No.1 (Jan-2023)

http://dx.doi.org/10.12785/ijcds/130124

Data Mining Techniques in Disease Classification: Descriptive Bibliometric Analysis and Visualization of Global Publications

Kaushalya Dissanayake[1], Md Gapar Md Johar[2] and Nishani H. Ubeysekara[3]

[1]*School of Graduate Studies, Management and Science University, Shah Alam, Malaysia*
[2]*Information Technology and Innovation Centre, Management and Science University, Shah Alam, Malaysia*
[3]*Ministry of Nutrition, Colombo, Sri Lanka.*

Received 11 Mar. 2022, Revised 14 Sep. 2022, Accepted 12 Jan. 2023, Published 31 Jan. 2023

Abstract: Many literature searches are required in scientific study, and these take a significant amount of time and effort. The bibliometric analysis is useful for locating research hotspots and gaining an understanding of research trends, according to the published literature. This bibliometric analysis was carried out with the assistance of the tools Bibliometric package in R (biblioshyny) and VOSviewer; the data used in this analysis was primarily derived from the Scopus repository to analyze the data mining approaches for disease classification (DMDC). A sample of 804 articles was selected by utilizing a query including essential key terms such as (”data mining” OR “data-mining”) AND (“disease classification” OR “disease identification” OR “disease prediction”). Overall, the findings of the study indicate the highest number of publications on applying DMDC published in 2019 with 141 research articles. As for individual researchers, the most productive authors are Jabbar MA, Li J, and Wang X. Jabbar MA in the field of DMDC, and 57 articles (1.2%) were written by one author, while the rest of the 747 articles were written by multiple authors. The Advances in Intelligent Systems and Computing journal (26 articles) has the greatest number of published articles connected to the DMDC field. Total of 804 articles in 474 distinct journals, there are 57 journals that have already published more than three papers, accounting for 12.03%. The USA was the most productive country, and the University of California is the affiliated university that comes from most top research in this field. The most cited research article published by Moore JH in 2006 included 489 citations. There are different types of diseases identified using data mining techniques such as heart disease, Brest cancer, liver disease, chronic kidney disease, Parkinson’s disease, diabetes mellitus, and Alzheimer’s disease. The most widely used algorithms in the research community include the random forest, decision tree, support vector machine, and naive bayes algorithms. In the future, the field of data mining could grow in many different directions and could be an effective way to increase the accuracy of disease prediction.

Keywords: Disease Classification, Data mining, Bibliometric analysis

1. INTRODUCTION

Using advanced statistical methods, data mining, sometimes referred to as Knowledge Discovery in Databases (KDD), is a promising and innovative method for discovering links and patterns in large databases [1][2][3]. Data mining is becoming an important research field because there has been a lot of growth in the types of data available. This has led to a lot of interest from both academia and industry.

Data investigation is the extraction of valuable information by handling the data that is a prime concern in recent years. Besides this concern, the identification of features is another problem in data analytics. Data mining methods are used to customize and monitor information systems to find information and key features in a dataset[3][4].

Data mining is a subfield of artificial intelligence that entails the examination of enormous amounts of data in order to uncover previously unrecognized patterns[5]. Additionally, data mining approaches can be used to group variables with similar behaviours and forecast future events.

In biomedical diagnosis, patients suffer from more than one kind of illness in the same form that can have redundant and interrelated symptoms and signs that doctors do not identify correctly[6]. Extracting useful and unique information from medical data is referred to as data mining in the context of health care[7][8][9]. Information systems in hospitals are becoming more refined as a result of the rapid growth of science and technology, and the number of healthcare data being stored is growing as well. The number of researchers in the healthcare data mining field is growing rapidly, and data mining applications in the medical sector is gaining popularity among academic researchers and developers[8][10].

E-mail address: kaushi.dula@gmail.com

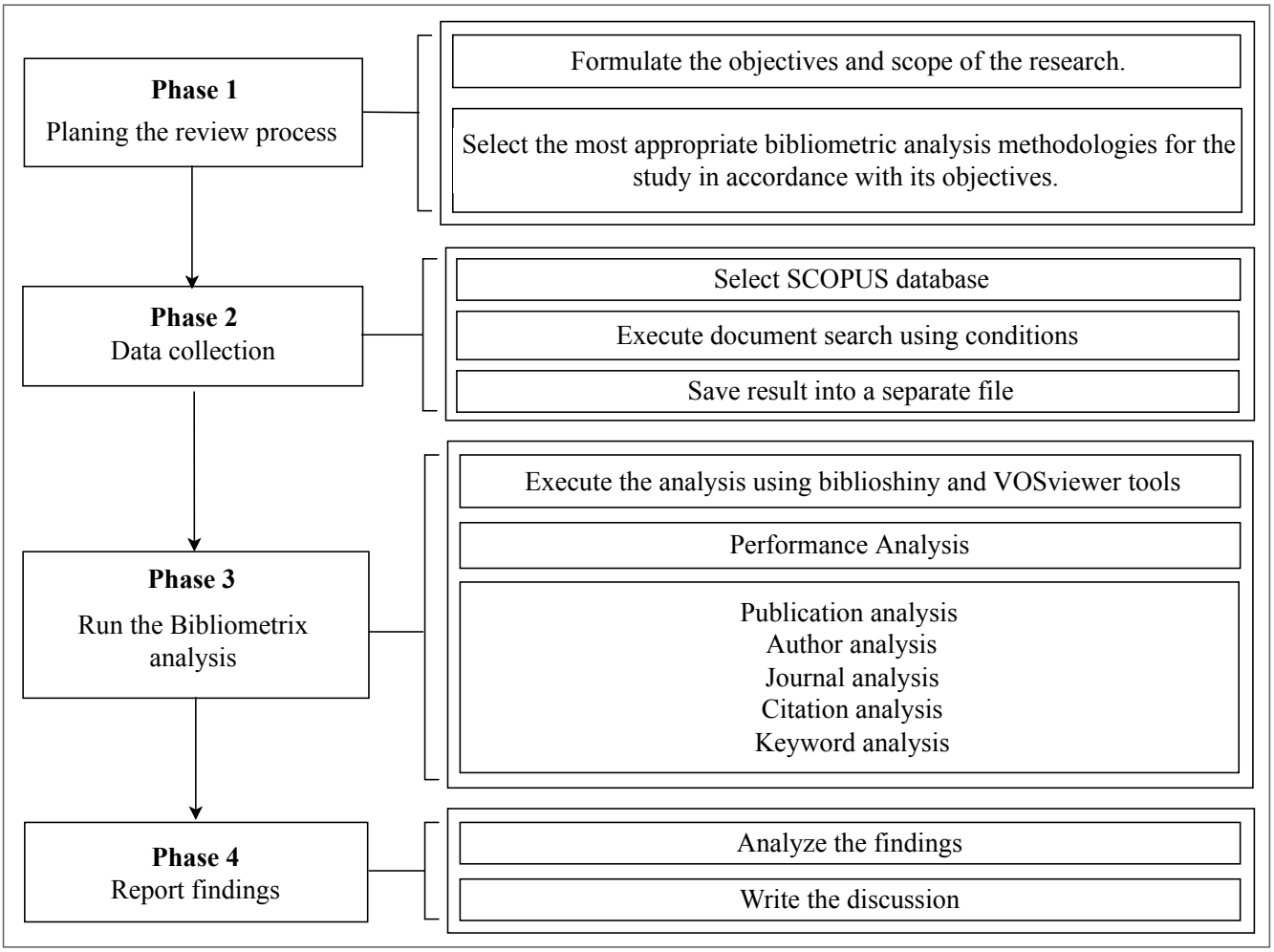

Figure 1. Framework for the bibliometric analysis. Note(s): This figure illustrates the phases and steps that are followed in each phase. The framework consists of four primary phases are as follows: planning the review process, data gathering, conducting the bibliometric analysis, and reporting findings.

Previous researchers have demonstrated that the employment data mining technologies in the healthcare industry consume less time to predict diseases with more reliable outcomes and reduce memory and processing resources[3][4][5][11]. Decision-making practices, such as diagnosis, prognosis, and medicine planning, can be built by applying data mining techniques. When assessed and confirmed, the accuracy level of practices could be embedded into clinical information systems. It has been used to solve a specific diagnostic problem for unspecialized doctors to make decisions accurately within less time.

Since substantial amounts of literature must always be searched when doing scientific research, it is critical to swiftly identify research hotspots and comprehend the primary development directions of current studies from those publications. Manual analyses, which rely significantly on time-consuming and labor-intensive methods but also on the researcher's personal experience and study interests, sometimes fail to adequately extract implicit information and underlying rules.

Nowadays, researchers are increasingly using statistical approaches to analyze the content of literature, citations, author information, external characteristics of documentation, and other relevant information to provide valuable research guidance. Research gaps, leading topics, and recent advancements in a particular field of study can all be visualized visually and geographically using bibliometric methods[12][13][14]. As a result, it has developed into an important instrument to evaluate national and worldwide productivity of research, citation analysis, global collaboration, developing advancement in the field, and the growth of specialized fields of research. While existing bibliometric techniques emphasize the use of statistical and mathematical techniques to define, assess, and forecast the state and

TABLE I. Criteria of search

Category of search	Criteria of search	No of articles
Key words	(”data mining” OR “data-mining”) AND (”disease classification” OR “disease identification” OR “disease prediction”)	886
Published year	2001 to 2021	877
Document type	Article, Review and Conference paper	816
Language	English	804

progression of technology and science. And they typically employ relatively simple mathematical techniques (which are essentially basic mathematics methods), derive relatively straightforward conclusions, and provide only limited insight for scientific research[15][16].

Numerous bibliometric software, such as VOSviewer, Bibliometric package in R (Biblioshiny), BibExcel, HistCite, Cite Spqace, Nvivo, Perish, Pajek, UCINET, Gephi, Leximancer, SciMat, and Sci2 have been developed to assist researchers in a variety of fields. Mainly used these software tools to identify research hotspots, evaluate the collaborative view on a subject, and develop maps knowledge[17].

It is the purpose of this research article to conduct a bibliometric analysis of research topics that are related to applying data mining techniques to disease classification (DMDC). The Scopus database has been used to study the literature from 2001 to 2021 in order to provide a comprehensive review of the important aspects of data mining-related disease classification publications[18].

Additionally, the clear, informative images offered in this work highlight research accomplishments in the DMDC domain, which can assist researchers and practitioners in identifying the underlying implications of authors, journals, countries, and trending research topics.

While this is not a comprehensive survey of the literature on DMDC, it highlights how bibliometric approaches may be used to identify underlying knowledge areas. The remainder of this article is organized in the following manner: the materials and methods outlined in section 2, followed by Section 3 which contained discussion and conclusion, and in section 4 mainly included limitations and future study.

2. Materials and Methods

The processes for doing bibliometric analysis, as well as the essential procedures that must be followed, are demonstrated in this section of the article. As shown in Figure 1 is the conceptual framework for the proposed bibliometric analysis. The phases of the proposed framework are described in detail in the following sections.

A. Phase 1: Planning the review process

The first stage is to define the bibliometric study's objectives and scope, which must occur prior to selecting bibliometric analytic tools and collecting bibliometric data.

1) Step 1: Formulate the objectives and scope of the research.

There were no published works at the time of the study for a comprehensive analysis of the field data mining applications for disease classification using bibliometric analysis. Data mining for disease classification has been the focus of this bibliometric study, which aims to provide an overview of the publications, most contributing researchers, productive journals, most cited countries, the keyword analysis and identify the keywords in the implications and future research[18].

B. Phase 2: Data collection.

Phase 2 consists of gathering data necessary for the bibliometric analysis approaches that have been selected.

1) Step 1: Select the database.

The SCOPUS database was used to acquire the literature data used in this investigation. Among the comprehensive bibliographic databases, SCOPUS provides access to a wide range of online sources, including huge databases of citation; however, not every article and journal is included in SCOPUS[18].

2) Step 2: Execute document search using conditions.

The purpose of this research is to evaluate the knowledge related to disease classification using data mining techniques. The keywords (”data mining” OR “data-mining”) AND (“disease classification” OR “disease identification” OR “disease prediction”) were used to conduct the investigation. All keywords are included in the query of search in the article abstract, title, and keywords throughout each document. The number of articles returned as a consequence of the search process is shown in Table I.

After the initial keyword search, 886 items were selected. It was important to screen the articles after the initial keyword search to identify those with the most relevance to the study's goals. Therefore, three steps were followed to search and retrieve the data. At the beginning of 1999, a literature search is performed. This study mainly focuses on the enhancement of the research field in 20 years. In the next step, nine articles were eliminated that published year does not contain in 2001 to 2021.

As a next searching criteria document type was considered and included research articles, reviews, and conference papers. Book chapters, conference review, notes,

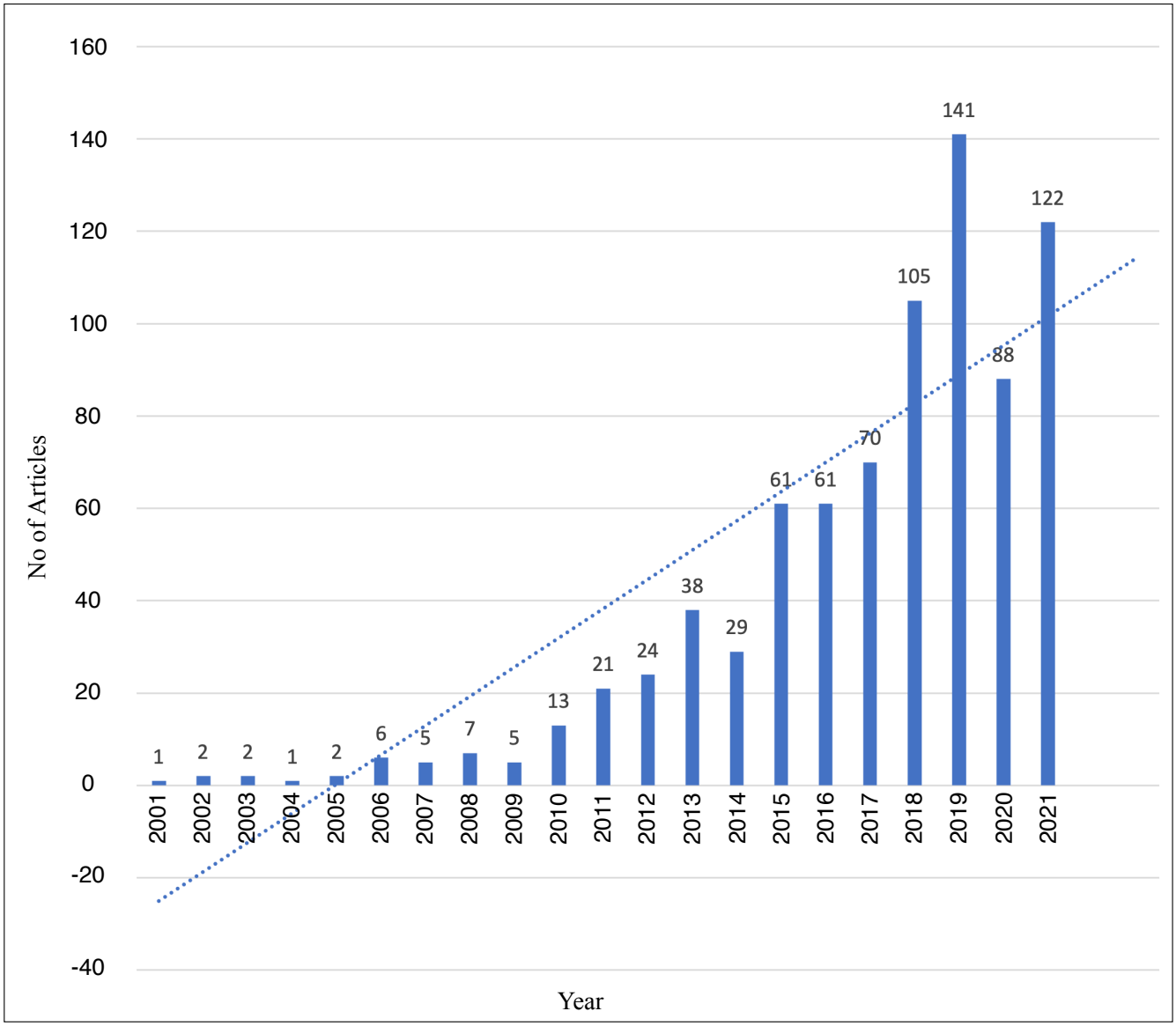

Figure 2. Growth of annual publication. Note(s): This graph represents the publication trend of the usage of data mining techniques for disease classification between 2001 and 2021. The data was retrieved from the Scopus database using keywords ("data mining" OR "data-mining") AND ("disease classification" OR "disease identification" OR "disease prediction"). The Linear forecast of the annual publications is represented on the plot using a dotted line.

book and retracted were excluded and 816 articles were retrieved covering disciplines of Medicine, Genetics and Molecular Biology, Biochemistry, Decision Sciences and Health Professions, Computer Science, Mathematics, and other related topics. In the final search, criteria apply the article published in the English language and selected 804 articles.

Based on the 804 papers that were selected, a scientometric analysis was done. The goal of this analysis was to get a complete picture of the research field and to give more detailed information than the ones that had been studied before.

3) Step 3: Export result data.

As part of the study, it was necessary to convert the final dataset, which comprised of 804 journal articles, into an accessible file format. The file format of the retrieved dataset was determining the grounded software that was used for analysis. Hence, information for publications was exported from the Scopus database into CSV format in Microsoft Excel. This included the authors of the document and the title of the source document, the number of citations, the countries of the authors and their affiliations, and the keywords and abstracts of the articles.

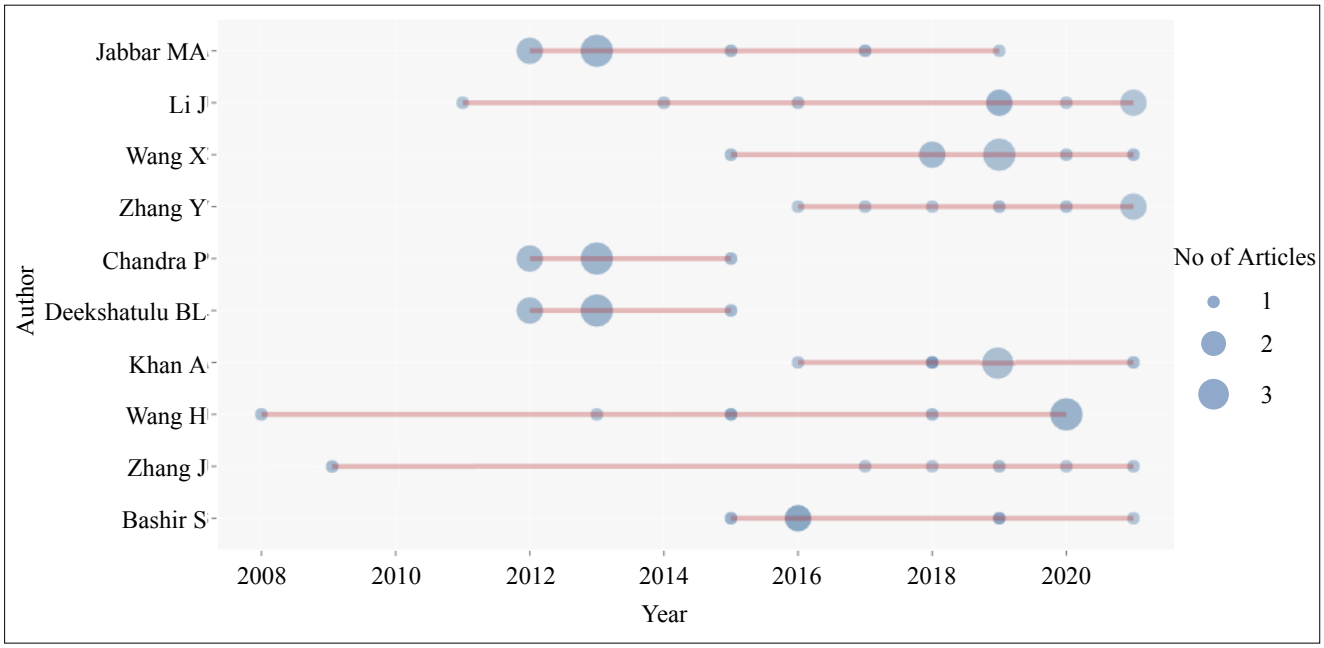

Figure 3. Authors' publications over time. Note(s): The figure illustrates the top 10 of the most productive authors in terms of publications on classifying diseases using data mining techniques. The horizontal line outlines the period that was active by each author made the publications. The number of articles published by each author each year is represented by the size of the circle.

C. Phase 3: Run the bibliometric analysis.

Both VOSviewer and Bibliometric software tools were utilized in this study for data analysis. VOSviewer was developed at the University of Leiden in the Netherlands by Ludo Waltman and Nees Jan van Eck. And also, Corrado Cuccurullo and Massimo Aria designed Bibliometric in Italy at the University of Naples.For this bibliometric analysis used version 1.6.17 of VOSviewer and version 3.1 of the Bibliometric tool[19][20]. The selection of these two software was made because both software is available to bibliometric researchers for free. The Bibliometric software is a collection of packages for conducting research related to the quantitative in the fields of scientometrics and bibliometrics. The VOSViewer, a software application for visualizing scientific data, was utilized to generate a network graph related to the bibliometric based on the co-occurrence of retrieved publications. Density maps and networks are visualization techniques that use colors, line thickness of connection, font sizes, and circles of varying sizes to depict various factors.

TABLE II. Type of retrieved documents.

Document type	Frequency	Proportion (%)
Research article	470	58.46
Conference paper	294	36.56
Review article	40	4.98
Total	804	100

D. Phase 4: Report results and findings.

This section illustrates and elucidates the results and findings of the descriptive and bibliometric analyses. A total of 804 documents were retrieved; 470 were research articles, 294 were conference papers, and the remaining 40 were categorized as review papers summarized in Table II.

1) Number of Publications analysis.

The quantity of publications is an essential indicator of the scientific research community's development tendencies. The progression of the article published annually was evaluated in the initial stage of research, as illustrated in Figure 2. The goal of this study is to figure out how researchers' interest in topics related to the DMDC changes over time.

In order to create the graph in Figure 2, we used the data obtained from bibliometric software, which took into account the papers taken from the Scopus database between 2001 and 2021. The year of the publishing is denoted on the x-axis and the number of published articles every year is marked on the y-axis in the graph presented in Figure 2.

An annual growth rate was used to show how many papers were written in a given year. A relative growth rate was used to show how many publications grew over a certain amount of time. The annual growth rate was 25.89%, according to the findings of the investigation. The first sub-period (2001-2009) marked the commencement of the research field, with the publication of 31 documents. The number of publications steadily increased until 2019 and 141 documents were published in 2019, which achieve the highest number of publications. The growth rate slightly declined in 2020, but over time in 2021, it was inclined.

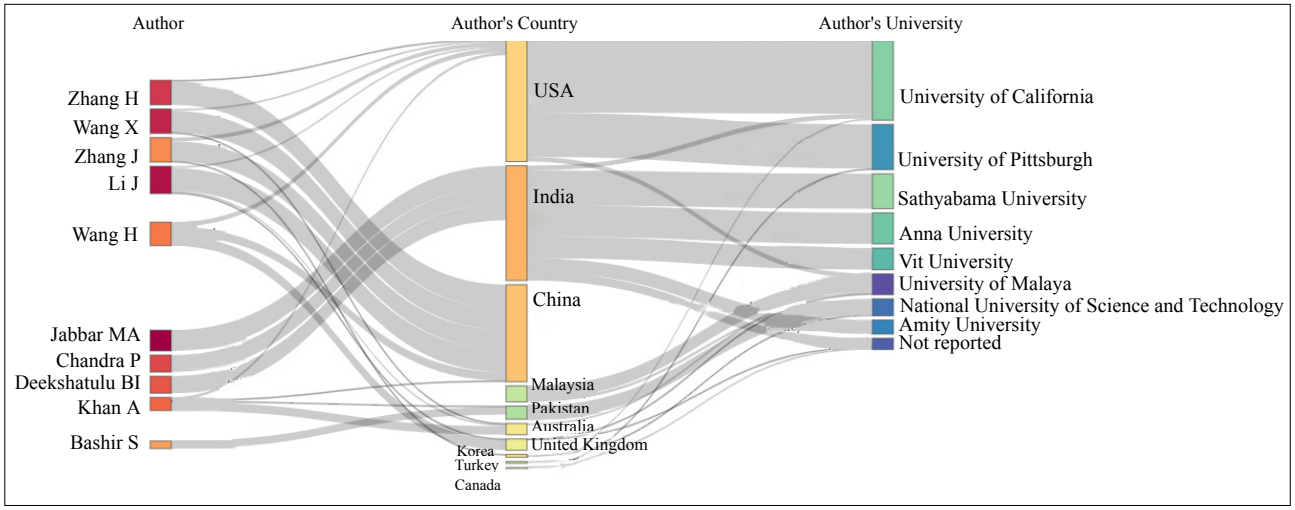

Figure 4. Three field plots of author, authors' country, and authors' university. Note(s): This figure illustrates the relationships between the most productive top ten authors in the left field, countries in the middle field, and affiliated universities in the right field period of 2001to 2021.The distinct color is used to represent each author, author's country, and author's university.

Over the previous several years, there has been an increased interest in employing data mining techniques to forecast disease, and this trend is expected to continue in the future. According to the linear forecast plot, 108 research articles will be published in the 2022 field of DMDC.

2) Author analysis.

The academic community's core authors serve as a key source of the internal strength of the research discipline, and researchers might locate possible partners by looking at these authors. There were 2793 authors who participated in the publishing of articles and 3.47 average authors for each paper inside the area of the study that was chosen.

TABLE III. Most productive ten authors in DMDC publications.

Rank	Authors	Articles
1	Jabbar MA	8
2	Li J	8
3	Wang X	8
4	Zhang Y	7
5	Wang H	6
6	Chandra P	6
7	Deekshatulu BL	6
8	Khan A	6
9	Wang W	6
10	Zhang J	5

Among the DMDC researchers throughout the world, the top ten most engaged core authors are listed in Table III. These authors have created effective collaborations with researchers from different countries.

As for individual researchers, the most productive authors are Jabbar MA, Li J and Wang X. Jabbar MA who is based at the Vardhaman College of Engineering, India, and studies Association Rule, Na¨ıve Bayse, Random Forest and data mining etc. Next was LI J, who scholar at College of Automation, Harbin Engineering University, China. He dedicated his studies Alzheimer Disease, Brain Network, Functional Connectivity, Mild Cognitive Impairment etc. Furthermore, Wang X also provide equal contribution for the research field who based at Department of Psychiatry, The First Affiliated Hospital of Harbin Medical University, China and studies field of Auxiliary Diagnosis System; Bayesian Network; Data Mining etc.

Figure 3 illustrates the top authors over the time. According to the figure, all the top 10 authors maximumly published 3 papers annually. Jabbar MA, Chandra P and Dekakatal BL published in 2013, Wang X and Khan A in 2019 and Wang X in 2020. The 57 articles (1.2%) were written by one author, while the rest of the 747 articles (98.8%) were written by multiple authors according to the data retrieved in the analysis.

3) Analysis of Geographical distribution of authors.

According to the data analysis, the authors are coming from 56 different countries worldwide related to the field of DMDC. Figure 4 illustrates the productive leading ten countries with 10 authors and 10 affiliated universities of authors. The USA was the most productive country, and the University of California is the affiliated university that comes from most top research in this field. From the middle field it is noticed that the India, China, and Malaysia also have the largest number of publications in DMDC and cover most research area. The bibliometric package was used to analyze collaboration internationally depending on the origin country.

TABLE IV. Criteria of search

Rank	Journal	Number of articles
1	Advances in Intelligent Systems and Computing	26
2	Artificial Intelligence in Medicine	15
2	International Journal of Recent Technology and Engineering	15
3	Computers In Biology and Medicine	14
3	Journal of Biomedical Informatics	14
4	Journal of Medical Systems	13
5	Journal of Advanced Research in Dynamical and Control Systems	10
6	International Journal of Applied Engineering Research	9
6	Pervasive health: Pervasive Computing Technologies for Healthcare	9
7	International Journal of Innovative Technology and Exploring Engineering	8

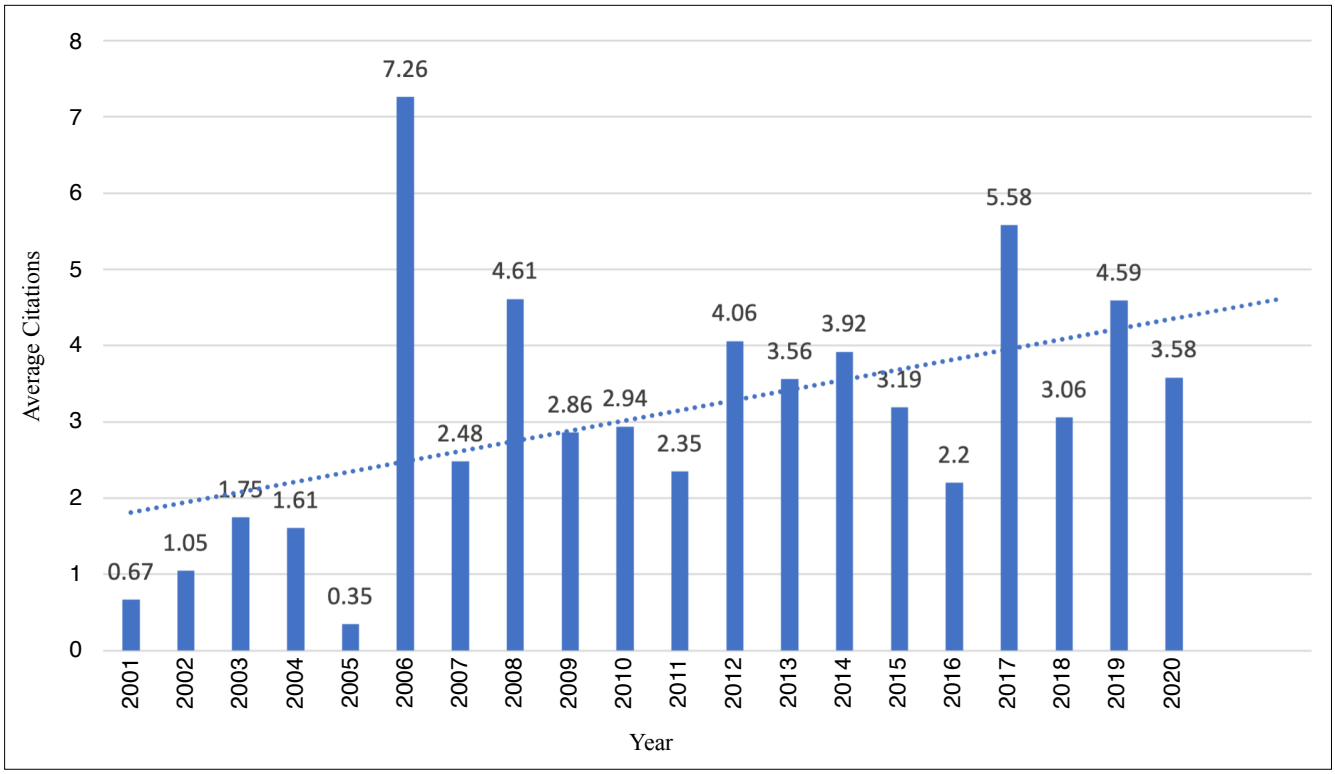

Figure 5. Average article citations per year. Note(s): This graph represents the publication trend of the article citations in the field of data mining techniques for disease identification. Each bar in a histogram represented the average amount of citations in each year. The Linear forecast of the average citations annually is depicted using a dotted line on the plot.

4) Journal analysis.

The researchers related to the field of DMDC selected to publish their articles in different journals around the world. At the time of executing the study, the data in Table 4 highlighted the topmost journals that are published the highest number of publications related to the research field.

While there are a total of 804 articles in 474 distinct journals, there are 57 journals that have already published more than three papers, accounting for 12.0%. According to the information shown in Table IV, the Advances in Intelligent Systems and Computing journal (26 articles) has the greatest number of published articles connected to the DMDC field. The second-place take two journals, that were published 15 articles in each journal, namely Artificial Intelligence in Medicine and International Journal of Recent Technology and Engineering. Computers In Biology And Medicine and Journal of Biomedical Informatics are got the same place in ranking with publishing 14 articles.

5) Average article citations analysis.

Citation trend analysis is carried out to determine the significance of a contribution in terms of the number of publications citing or referring to it.

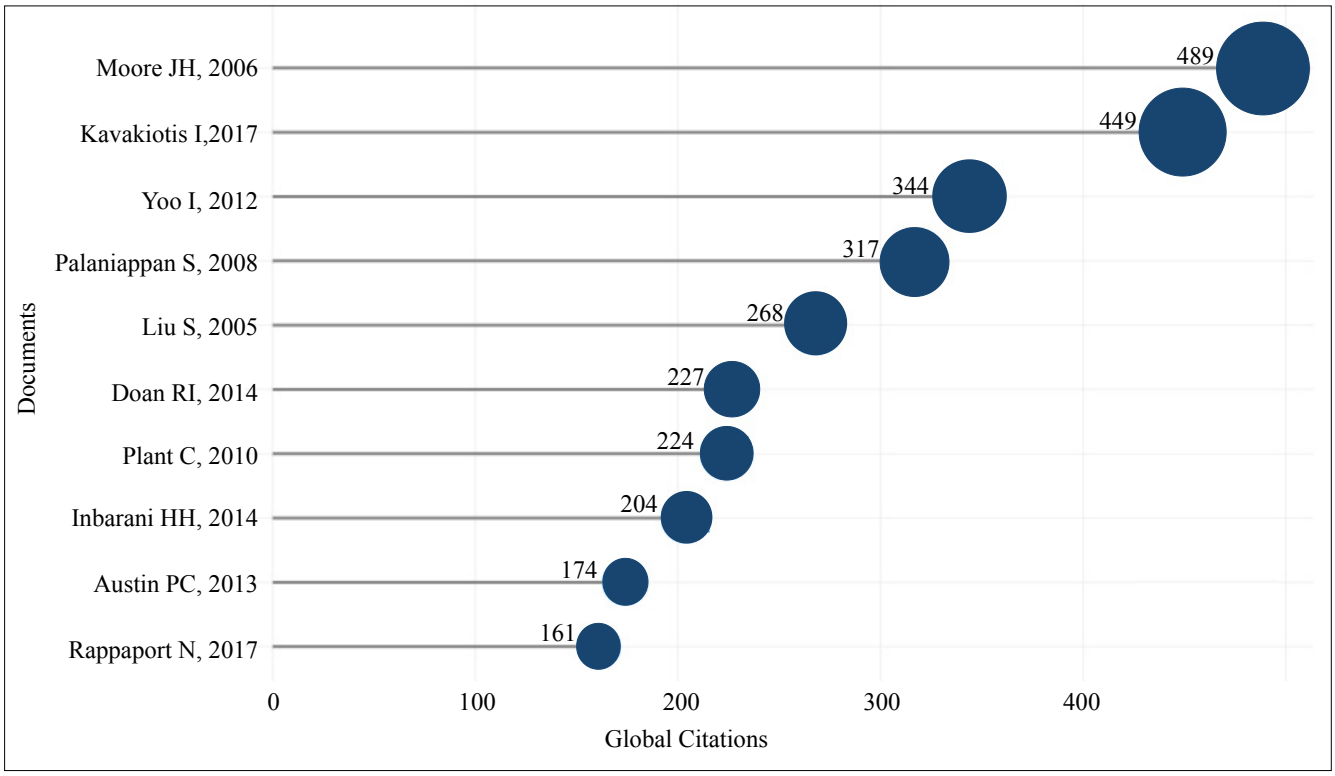

Figure 6. Top ten most cited articles of the DMDC field. Note(s): Most frequently cited articles among the scientific contributions to DMDC are depicted in this figure. The articles are sorted by the number of global citations. At the end of each circle with different sizes represents the number of citations obtained by each publication.

Figure 5 represents the average article citations per year of research constituent. According to the statistics in the figure average article citation fluctuated over the period of analysis. In 2006, the rapidly inclined take related to the average citation that is 7.26 per year. And also in 2008, 2010, 2012, 2014, 2017, and 2019 slight improvement in the average citation. Based on the linear forecast plot for average citations, more than 4.5 citations will be obtained in 2022.

6) Most cited papers analysis.

One another factor that can be used to determine the research quality is an analysis of citation. The Journal of Theoretical Biology published the article that received the most citations in the study field. The top ten cited articles in the field of DMDC are depicted in Figure 6.

The most cited research article published by Moore JH in 2006; entitled to research title "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility" archive citation counts 489. All the top ten cited publications achieved more than 150 citations.

India (2235) was the major place among cited countries, the next two places are taken by the USA (1443) and China (1060), according to the result of citation analysis.

7) Co-keyword analysis.

The term "keywords" indicate phrases or nouns that describe the major contents of an article. The analysis of co-keyword analysis was carried out in the next step of the study. The network map that is generated as a result of the keyword co-occurrence analysis highlights the most frequently encountered keywords, groups the closely related words, and shows connections to the other keywords. Consequently, it is possible to identify the most frequently used keywords by authors, as well as trends and patterns in the study field.

The keywords co-occurrence map was created using 1970 keywords extracted from the articles shown in Figure 7. The threshold of co-occurrence for terms is set to five, and the most used 15 terms and their link with other keywords are visualized using the VOSviewer to show the hotspot in the DMDC research field. The visualization items in the network are represented by their circles and label. According to object weight, the circle and label size are calculated. When the weight of an object is high, the circle and label connected with it are larger. Links between keywords are shown by the lines connecting them, and the colour of each item depends on the cluster to which it belongs. And also mostly related terms are classified into one cluster and depicted using one colour. The main five

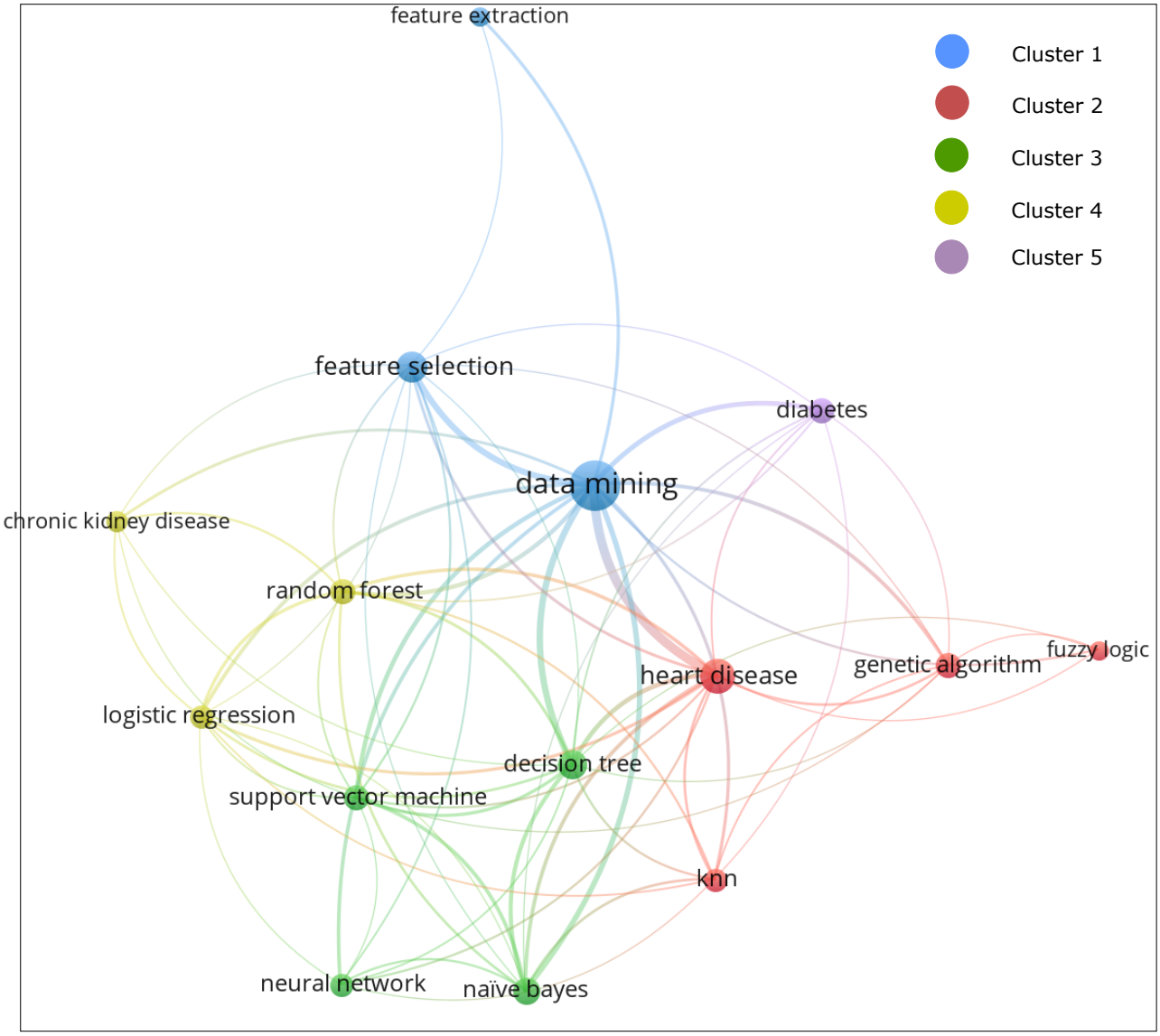

Figure 7. The keywords co-occurrence network map. Note(s): All keywords are categorized into five major clusters that are closely related keywords represented with five different colors. Each node in a network represents a keyword. The occurrence of the keyword is represented by the size of the node. The link between the nodes illustrates the co-occurrence between keywords. The thickness of the link signals the occurrence of co-occurrences between keywords.

clusters are represented using five colours.

As expected, the keyword “data mining” occupies the center of the map and accumulates 74 strengths of the total link and appears 45 times.

The map highlighted other keywords that have the highest influence such as classification (108), machine learning (103), heart disease (83), and feature selection (51). Figure 7 depicted the network of keywords belonging to five unique main clusters, each of the clusters corresponds to a subfield of DMDC. Based on the core documents, citations, centrality, h-index, and density, quantify the theme of each cluster. The data in Table 5 indicate top ten keywords were used in DMDC publications.

The evolution of different aspects related to the field of DMDC can be revealed through keyword analysis. There are different types of diseases identified using data mining techniques such as heart disease, Brest cancer, liver dis-

ease, chronic kidney disease, Parkinson's disease, diabetes mellitus, and Alzheimer's disease. According to the cluster represented by blue colour evident that feature extraction and feature selection were used when classifying the disease using data mining applications.

TABLE V. Top ten keyword in DMDC publications.

Rank	Keywords	Frequency	Total link strength
1	Data mining	297	589
2	Classification	108	263
3	Machine learning	103	222
4	Heart disease	83	208
5	Feature selection	51	113
6	Prediction	44	113
7	Decision tree	43	117
8	Naïve bayes	28	89
9	Support Vector Machine	25	64
10	Random forest	22	82

8) Trend Topic Analysis.

It is necessary to analyse the shifting theme evolution over time to discover developing and saturated issues. Additionally, by analysing the terms used by authors to represent their area of expertise, Bibliometric software was used to construct a trend topic analysis, as illustrated in Figure 8. The trend topic plot, which was made by looking at the main keywords utilized by the authors of the publications, shows how the use of DMDC has changed over time from 2009 to 2021.

The trend topic graph is generated using two main parameters; The value five is assigned as the lowest marginal frequency value of the word and the number of words in every year. Under these settings, as illustrated in Figure 8, the most frequently used terms for every year might be identified. The lines reflect the years in which that term was most widely used, and the median year represented using the bubble. The term frequency of median year is depicted top of the bubble.

The keyword "data mining" is the most used, appearing 299 times in 2018. Additionally, it's fascinating to study how the pattern developed. In 2021, most of the research will focus on "cardiovascular disease". In 2018, the topic trend is comparatively high, "data mining" was researched more, followed by words such as "classification", "feature selection" and "decision tree".

9) Topic Dendrogram.

The keywords conceptual framework that used, the substance of the major themes studied in the DMDP field can be expanded even more, as shown in Figure 9.

Topics dendrogram mainly formed with two clusters represented using two colors on the diagram. The initial cluster included with "Support vector machine", "K-means" and "Artificial Neural Network" can be categorized under most used data mining techniques. The second cluster is mainly split into two clusters categorized under various themes that contribute to data mining and diseases.

There is a broad agreement among the conclusions drawn from the prior analysis; thus, for different disease classify using different types of data mining algorithms.

3. Discussion and Conclusion

The DMDC knowledge map was produced using the information visualization tools bibliometric package in R (biblioshyny) and VOSviewer, using Scopusindexed literature from 2001 to 2021. The scientific footprint of data mining techniques used to identify various diseases is developing. The findings of this bibliometric study in the field demonstrate that the publications analysis, author analysis, journal analysis, citation analysis, and topical focus in the DMDC Field of research have been identified.

Researchers are becoming increasingly interested in studying how data mining might be used to identify disease, which is in line with the pressing need to identify diseases, support physicians, improve public health, and provide support to patients.

The DMDC knowledge map was produced using the information visualization tools bibliometric package in R (biblioshyny) and VOSviewer, using Scopusindexed literature from 2001 to 2021. The scientific footprint of data mining techniques used to identify various diseases is developing. The findings of this bibliometric study in the field demonstrate that the publications analysis, author analysis, journal analysis, citation analysis, and topical focus in the DMDC Field of research have been identified.

Researchers are becoming increasingly interested in studying how data mining might be used to identify disease, which is in line with the pressing need to identify diseases, support physicians, improve public health, and provide support to patients.

Researchers and clinicians can use the data to make more precise diagnoses of disease, which helps to establish future research goals and advances in this rapidly evolving field. Periodic bibliometric analysis, perhaps at 2009-2021 intervals, will let us see how this field changes over time. This will help us see how this field changes over time.

Several noteworthy findings of DMDC publications might be summarized as follows. There has been a steady rise in published and cited articles over the last few decades, the annual growth rate was 25.89%, according to the data in the Scopus database. 2019 had the most publications, 147 (18.2%), owing to the development of disease classification applications using data mining techniques in the medical industry.

Second, the Advances in Intelligent Systems and Computing, a peer-reviewed journal, has published the 26 (3.2%)

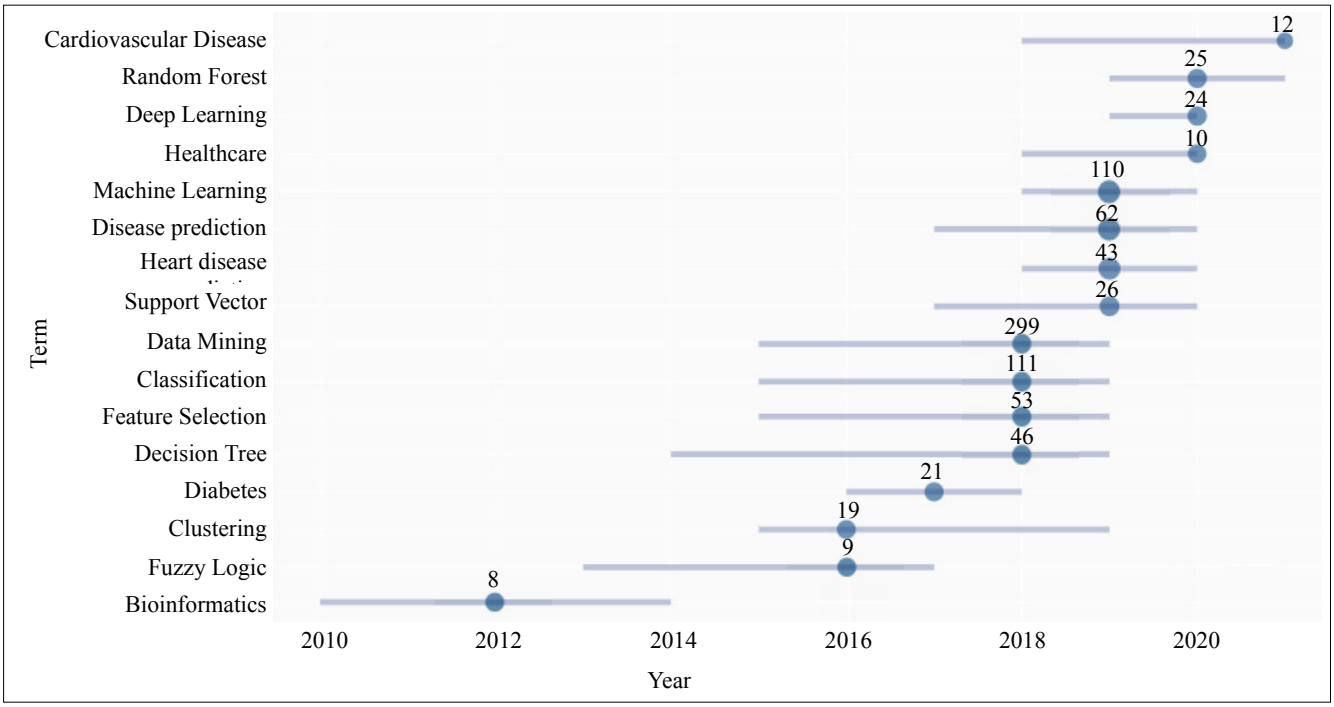

Figure 8. The trend topic plot. Note(s): The most relevant trend topics in publications on DMDC in terms of the total number of appearances in the period 2000–2021 are denoted in this figure. The horizontal line shows the most trending time period of each topic. The bubble indicates the median year of the topic that appeared.

the maximum number of articles related to the field of research. Three researchers, in particular, make a significant contribution to this research field, notably Jabbar MA, Li J, and Wang X, each of whom has published 8 articles in this field.

The frequency with which an article is cited as a referral in another publication reflects the publication's significance scientifically. In addition, the distribution of the most important publications in the subject of data mining can be identified through citation analysis. Thirdly, the average number of citations per year fluctuated throughout time, reaching a peak in 2006 obtained 7.2 citation per year. The results of the citation analysis highlighted, India is the most cited country (citations: 2235), followed by the United States of America (citations: 1443) and China (citations: 1060).

Fourth, a keyword is a symbol of a very active area of research. The keywords that receive the highest attention from the relevant scientific community, demonstrate the active areas in the research domain. We discovered several trends in the new studies, such as obtaining information from data mining utilizing various techniques based on the trend topic analysis and keyword network. Furthermore, the identification of diseases through the use of computational models is an unavoidable trend in future medical progress, and there have been several disputes concerning the accuracy of diagnosis in the past.

Using data mining algorithms to construct disease classification systems, which can be employed in a variety of situations such as emergencies, to reduce human errors and a shortage of experienced physicians are all possibilities.

A number of algorithms are used in diagnostics, including naive Bayes, support vector machines (SVM), logistic regression, k- means, decision trees, k nearest neighbour, genetic algorithms, and artificial neural networks. The most widely used algorithms in the research community include the random forest, decision tree, support vector machine and naive bayes algorithms. Each data mining algorithm, on the other hand, has its own set of pros and cons.

Several common strengths of data mining approaches may be found, such as appropriate computational accuracy and the capacity to deal with complex interactions between distinct features. The algorithms used in data mining are capable of extracting important knowledge from raw datasets; yet the models used in data mining are too sophisticated for human experts to understand and interpret, particularly in the case of black-box phenomena.

4. Limitations and Future Work

Despite the fact that we've examined the DMDC papers and shown the research accomplishments and potential impact according to the journals, research area, authors and countries, we still believe there's more to be done. According to the findings of the study, various limitations have been observed

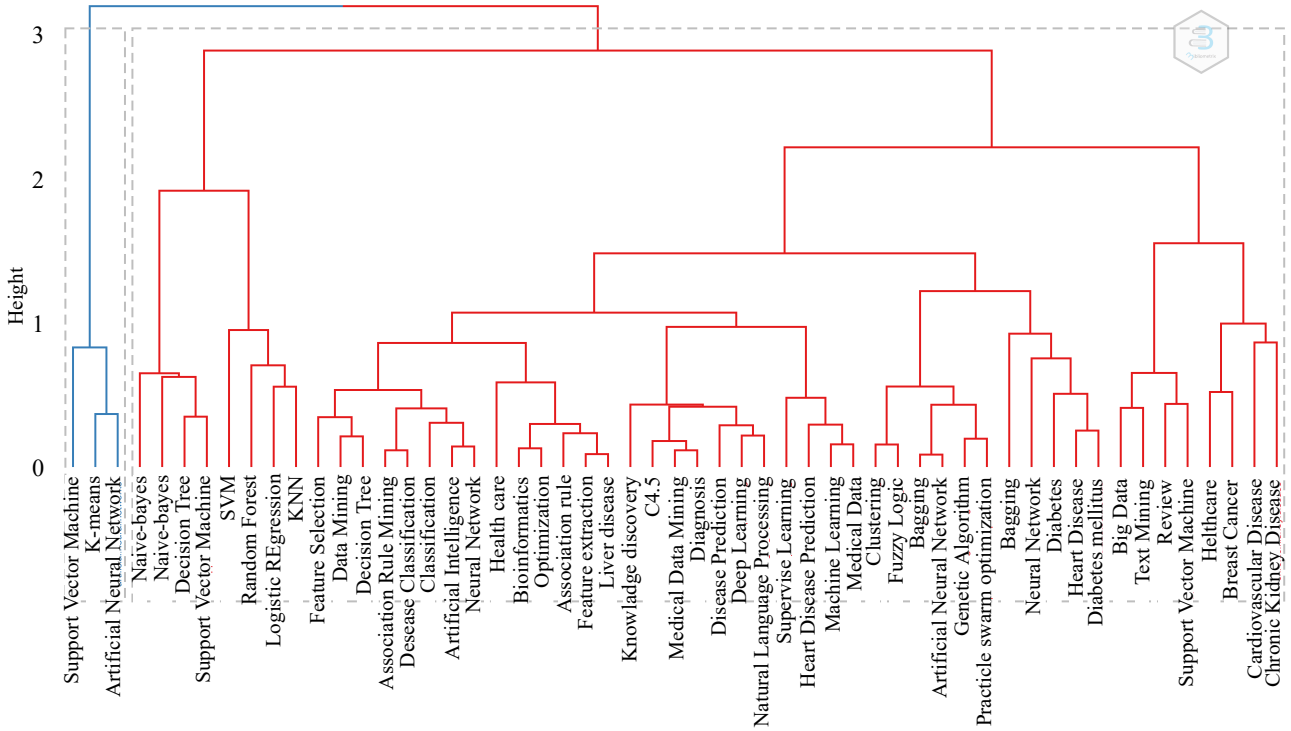

Figure 9. The topic dendrogram. Note(s): The following parameters were used to build the topic dendrogram: the number of terms, field author keywords were limited to 50 and multiple correspondence analysis, and the program was permitted to choose the number of clusters autonomously. The major two clusters are depicted using two separate colours on the diagram.

The data for this bibliometric analysis was obtained from the Scopus database. Improved and larger datasets are required to better understand research related to the field of DMDC. Additionally, we limit our analysis to articles from 2001 to 2021 that reference DMDC in the database and are publicly accessible on the internet. The authors acknowledge that they may have missed some significant research papers on DMDC and that certain pertinent records may have been omitted if the query terms used for subject searches did not match certain records. In the future, we hope to combine data from several databases to gain a better understanding of the area, researchers, and publications covering the longer period.

References

[1] A. K. Verma, S. Pal, and S. Kumar, "Classification of skin disease using ensemble data mining techniques," *Asian Pacific journal of cancer prevention: APJCP*, vol. 20, no. 6, p. 1887, 2019.

[2] P. Singh and N. Singh, "Role of data mining techniques in bioinformatics," *International Journal of Applied Research in Bioinformatics (IJARB)*, vol. 11, no. 1, pp. 51–60, 2021.

[3] J. A. Moral-Muñoz, E. Herrera-Viedma, A. Santisteban-Espejo, and M. J. Cobo, "Software tools for conducting bibliometric analysis in science: An up-to-date review," *Profesional de la Información*, vol. 29, no. 1, 2020.

[4] A. F. Choudhri, A. Siddiqui, N. R. Khan, and H. L. Cohen, "Understanding bibliometric parameters and analysis," *Informatics*, 2015.

[5] S. Joshi and M. K. Nair, "Prediction of heart disease using classification based data mining techniques," in *Computational Intelligence in Data Mining-Volume 2*. Springer, 2015, pp. 503–511.

[6] M. L. Kolling, L. B. Furstenau, M. K. Sott, B. Rabaioli, P. H. Ulmi, N. L. Bragazzi, and L. P. C. Tedesco, "Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development," *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3099, 2021.

[7] E. Parva, R. Boostani, Z. Ghahramani, and S. Paydar, "The necessity of data mining in clinical emergency medicine; a narrative review of the current literatrue," *Bulletin of Emergency & Trauma*, vol. 5, no. 2, p. 90, 2017.

[8] E. Shirzad, G. Ataei, and H. Saadatfar, "Applications of data mining in healthcare area: A survey," *Engineering and Applied Science Research*, vol. 48, no. 3, pp. 314–323, 2021.

[9] K. Dissanayake and M. G. Md Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, 2021.

[10] D. Mullaivanan and R. Kalpana, "A comprehensive survey of data mining techniques in disease prediction," *Challenges and Applications of Data Analytics in Social Perspectives*, pp. 27–53, 2021.

[11] A. Suragala, P. Venkateswarlu, and M. China Raju, "A comparative study of performance metrics of data mining algorithms on medical data," in *ICCCE 2020*. Springer, 2021, pp. 1549–1556.

[12] G. Manyangu, B. Dineen, R. Geoghegan, and G. Flaherty, "Descriptive bibliometric analysis of global publications in lifestyle-based preventive cardiology," *European Journal of Preventive Cardiology*, vol. 28, no. 12, pp. 1303–1314, 2021.

[13] B. S. dos Santos, M. T. A. Steiner, A. T. Fenerich, and R. H. P. Lima, "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018," *Computers & Industrial Engineering*, vol. 138, p. 106120, 2019.

[14] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *Journal of Business Research*, vol. 133, pp. 285–296, 2021.

[15] Y. Hu, Z. Yu, X. Cheng, Y. Luo, and C. Wen, "A bibliometric analysis and visualization of medical data mining research," *Medicine*, vol. 99, no. 22, 2020.

[16] D. Maier, "Building materials made of wood waste a solution to achieve the sustainable development goals," *Materials*, vol. 14, no. 24, p. 7638, 2021.

[17] S. A. Abd Karim and P. N. Nohuddin, "Bibliometric analysis of data mining on medical imaging," in *Journal of Physics: Conference Series*, vol. 1997, no. 1. IOP Publishing, 2021, p. 012017.

[18] J. Z. Zhang, P. R. Srivastava, D. Sharma, and P. Eachempati, "Big data analytics and machine learning: A retrospective overview and bibliometric analysis," *Expert Systems with Applications*, vol. 184, p. 115561, 2021.

[19] M. Aria and C. Cuccurullo, "bibliometrix: An r-tool for comprehensive science mapping analysis," *Journal of informetrics*, vol. 11, no. 4, pp. 959–975, 2017.

[20] H. Derviş, "Bibliometric analysis using bibliometrix an r package," *Journal of Scientometric Research*, vol. 8, no. 3, pp. 156–160, 2019.

Kaushalya Dissanayake She is currently PhD research scholar in Computer Science, School of Graduate Studies, Management and Science University, Malaysia. She received MSc in Information Technology (Specialized Cyber Security) from Sri Lanka Institute of Information Technology, Sri Lanka. She completed her BSc from the University of Ruhuna, Sri Lanka. Her research interests are in the field of Machine Learning, Data mining, Artificial Inteligence and Deep learning.

Md Gapar Md Johar He is Senior Vice President Research, Innovation, Technology and System of Management and Science University, Malaysia. He is a professor in Software Engineering. He holds PhD in Computer Science, MSc in Data Engineering and BSc (Hons) in Computer Science and Certified E-Commerce Consultant. He has more than 40 years of working and teaching experience in various organizations include Ministry of Finance, Ministry of Public Enterprise, Public Service Department, Glaxo Malaysia Sdn Bhd and Cosmopoint Institute of Technology. His research interests include learning content management system, knowledge management system, blended assessment system, data mining, RFID, e-commerce, image processing, character recognition, data analytics, artificial intelligent and healthcare management system.

Nishani H. Ubeysekara She is a consultant in Community Medicine affiliated to the Ministry of Health, Sri Lanka. She has got her MBBS degree from University of Sri Jayawardenepura, Sri Lanka and MSc and MD degrees in Community Medicine from University of Colombo, Sri Lanka. She has started her carrier as a medical doctor and has served Ministry of Health Sri Lanka for 22 years in curative sector and in preventive sector. She is a consultant in public health and has national and international experience in disease prevention, epidemiology and health promotion specially in prevention of Non Communicable diseases. She was a undergraduate and postgraduate medical trainer attached to the Faculty of Medicine , University of Ruhuna, Sri Lanka and also a trainer of public health staff at provincial and national level. She was a member of Endocrine and Immunology research team at medical school of Cardiff University , Wales and an international member of Faculty of Public Health, Royal College of Physicians, UK. At present she is the deputy Director of Teaching Hospital, Karapitiya, Galle, Sri Lanka.