January 2021

Bibliometric of Feature Selection Using Optimization Techniques in Healthcare using Scopus and Web of Science Databases

Rahul Joshi
Symbiosis International University, rahulj@sitpune.edu.in

Harita Gadikta
Symbiosis International University, gadikota.harita@sitpune.edu.in

Saneeka Kharat
Symbiosis International University, kharat.saneeka@sitpune.edu.in

Soumi Mandal
Symbiosis International University, soumi.mandal@sitpune.edu.in

Kalyani Kadam
Symbiosis International University, kalyanik@sitpune.edu.in

See next page for additional authors

Authors
Rahul Joshi, Harita Gadikta, Saneeka Kharat, Soumi Mandal, Kalyani Kadam, Anupkumar M. Bongale Dr., and Siddhant Pandit

Bibliometric of Feature Selection Using Optimization Techniques in Healthcare using Scopus and Web of Science Databases

Harita Gadikota, Saneeka Kharat, Soumi Mandal, Siddhant Pandit, Rahul Joshi, Kalyani Kadam, Dr. Anupkumar M. Bongale

Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India

Email id: rahulj@sitpune.edu.in

Abstract

Feature selection technique is an important step in the prediction and classification process, primarily in data mining related aspects or related to medical field. Feature selection is immersive with the errand of choosing a subset of applicable features that could be utilized in developing a prototype. Medical datasets are huge in size; hence some effective optimization techniques are required to produce accurate results. Optimization algorithms are a critical function in medical data mining particularly in identifying diseases since it offers excellent effectiveness in minimum computational expense and time. The classification algorithms also produce superior outcomes when an objective function is built using the feature selection algorithm. The solitary motive of the research paper analysis is to comprehend the reach and utility of optimization algorithms such as the Genetic Algorithm (GA), the Particle Swarm Optimization (PSO) and the Ant Colony Optimization (ACO) in the field of Health care.

The aim is to bring efficiency and maximum optimization in the health care sector using the vast information that is already available related to these fields. With the help of data sets that are available in the health care analysis, our focus is to extract the most important features using optimization techniques and work on different algorithms so as to get the most optimized result.

Precision largely depends on usefulness of features that are taken into consideration along with finding useful patterns in those features to characterize the main problem. The Performance of the optimized algorithm finds the overall optimum with less function evaluation. The principle target of this examination is to optimize feature selection technique to bring an optimized and efficient model to cater to various health issues.

In this research paper, to do bibliometric analysis Scopus and Web of Science databases are used. This bibliometric analysis considers important keywords, datasets, significance of the considered research papers. It also gives details about types, sources of publications, yearly publication trends, significant countries from Scopus and Web of Science. Also, it captures details about co-appearing keywords, authors, source titles through networked diagrams. In a way, this research paper can be useful to researchers who want to contribute in the area of feature selection and optimization in healthcare. From this research paper it is observed that there is a lot scope for research for the considered research area. This kind of research will also be helpful for analyzing pandemic scenarios like COVID-19.

Keywords: Bibliometric, feature selection, optimization techniques, healthcare, healthcare datasets etc.

1. Introduction

The immense growth in the field of technology could be a boon to various fields. Health care is one of the most important domains that could benefit with the growth in technology [1-2]. Technology could transform unsustainable systems into sustainable ones, provide effective solutions for diseases and predict diseases with provided features [4-5]. The field of health care has a lot of real time data available for various diseases which could efficiently be helpful if studied well and used with technology to further deduce patterns [3]. Improving the quality of life is a major benefit of integrating new innovations into health care. One of the fundamental challenges in health care is to achieve an efficient utilization of the growing technology [6-9]. There has been an increase in the amounts of data that is being gathered in health care and medical systems and the convergence of various domains are leading healthcare as well as medical research to a new direction of precision. The trends that are captured bring a unique and good opportunity in solving various critical tasks in healthcare research [7]. However, such precision largely depends on if useful features are taken into considerations along with finding useful patterns in these features to characterize the main problem [10].

There have been many efforts made using different approaches in this department. Our research extends along the lines of machine learning and optimization methods on feature selection to build novel models for solving and predicting medical challenges in health care [11-13]. The aim is to bring an optimized and efficient model to cater to various health problems.

The data pre-processing step is a significant advance in data mining, since the quality of choices depends on the nature of information. Upgrading the healthcare database improves the heart of medical analysis [14]. Data pre- preparing steps are data integration, data cleaning, data reduction and data transformation that is feature subset selection [15-20]. A few ascribes of datasets might be repetitive as their data might be accommodated in other attributes. Some additional traits can build computation time which in turns affects in determining accuracy. Some information in the dataset may not be valuable for determination and can be dispensed with prior to learning.

The objective of feature selection is to locate a base arrangement of traits with the end goal that the subsequent likelihood appropriation of the information category is as close as possible to the native distribution obtained using all traits [21-25]. Data reduction lessens the quantity of highlights, and eliminates immaterial, excess, or noisy information. This reduction has extraordinary quick impacts on accelerating data mining calculation, and bettering mining execution, for example, predictive precision and output comprehensibility [26-27].

In this research we present use of hybrid approach, which uses optimization algorithms such PSO, K-Means, Ant Optimization with classifiers used in the classification process. The resulted subset of features is provided as input to classifiers [28].

The goal is to achieve maximum efficiency. After Data Pre-Processing steps, Cluster Optimization Technique is used for clustering the individuals into subdivided groups [29]. The higher the values more desirable is the solution for it. Incremental Clustering algorithms are required, which can be used in the healthcare datasets for getting the optimized and characterized output [30-32]. Incremental clustering algorithms are further subdivided as K-means clustering, DBSCAN, Mean Shifting Algorithm and centroid based algorithm [33-35]. Less computation is observed in K-means clustering but the drawback is it takes random points which causes various results. DBSCAN (Density Based Spatial Clustering Application) is showing best results as it is having dynamic database. So without rerunning or restarting the new clusters are formed [36-38]. The clustering analysis can be observed by hard and soft clustering along with the priorities in it. Based on the count of clusters the priorities are determined. Probability of data points are assigned to the clusters known as soft clustering [39-40]. Randomly each data pint accessing the clusters are hard clusters. Evaluation of the clusters are quite important to get the high level output. Some key features like the scalability, high dimension ability, interpretable and reusable for the clustering techniques [41]. Evaluation and Checking in terms noisy datasets, erroneous dataset or missing datasets. These type of data can cause Poor quality cluster; hence the algorithms must be sensible. Validating the optimized clusters require internal and external clustering evaluation [42]. Using the internal data as well as the only external data can interpret the validation. But the most reliable and adaptable for the datasets are using Relative Clustering Validation, it helps to evaluate the varying parameters and find the optimal solution [42]. Optimized classifier shows the high accuracy to determine the optimized selected features. Optimization algorithms are the Genetic Algorithms(GA), the Swarm Intelligence (SI) techniques, the Particle Swarm Optimization(PSO) and the Ant Colony Optimization (ACO). Table 1 shows summary of keywords, datasets and significance related to considered research papers for the undertaken research.

Table 1: Summary of Referred Papers – Keywords, Datasets, Significance

Reference Number	Year	Keywords	Datasets	Significance
1	2020	Energy efficiency, IoT device, Smart healthcare, Sustainability, Particle swarm optimization(PSO)	Systematic student perspective healthcare data produced utilizing UCI dataset	Ensured the goodness of the EEPSOC-ANN model
2	2020	COVID-19, DNA-encoded chemical library, Focused DECL, Protease inhibitor, SARS-CoV-2	COVID-19 related research activities	COVID-19 affects DNA of the patients
3	2020	Classification, extraction, Heart disease prediction, Machine learning	Heart disease dataset	Data-driven methodologies are found reasonable for predicting heart disease
4	2020	Automated diagnosis, Expert health systems, PPG derivatives, Feature extraction, Myocardial infarction	Myocardial infarction (MI)	SVM the linear classification technique, positive productivity and detection accuracy of

		(MI), Variability analysis, Photo plethysmography (PPG)		98.10%,92.70% and 95.40%
5	2020	Autism Spectrum Disorder, Cultural Algorithm, Data Mining, Feature Selection, Optimization Techniques, Symmetrical Uncertainty	Autism colossal datasets	Regulating the high pertinent and low redundant features from the dataset using feature selection and optimization
6	2020	Affective computing, Healthcare computing, Emotion identification, Genetic algorithms, Optimization tasks and Feature Selection	AMIGOS dataset -A Dataset for Personality, Mood Research on Individuals, Affect, and Groups)	Sinusoidal chaotic maps interpretation, perform better and show better result compared to other maps like GA performance.
7	2020	Classification, Decision support system, Deep learning, Feature Selection, Healthcare Heart disease, Heart failure, Machine learning, Neural network, Weight optimization	Z-Alizadeh Sani heart disease dataset	formed a subset of special features which are selected by specific the techniques
8	2020	Kidney disease that is CKD (Chronic kidney disease), Fitness function, ITLBO (Improved Teacher Learner Based Optimization algorithm), Objective function.	CKD that is Chronic kidney disease	The given algorithm aims to deduce the total number of features required for diagnosing the Chronic kidney disease
9	2020	AI in healthcare, AutoML, Binary classification, Boosting, Cardiac	AI in healthcare related wok	Hyper parameter optimization techniques are executed

		arrest, Coronary heart disease, Early detection, Ensemble technique, Evolutionary algorithms, Extreme gradient boosting, Feature selection, Genetic algorithm, Heart attack, Hyper parameter tuning, Machine learning, Optimization, Optimized pipeline, Random forest, SMOTE, TPOT		to further improve the predictive model's performance
10	2020	Feature selection algorithm, ant colony optimization, classification, machine learning, Support vector machines and Parkinson's disease recognition	Some real-world dataset like Parkinson's disease	Accuracy for selected features- 99.50%
11	2019	Diabetes diagnosis, Feature selection, K-means algorithms, Meta-heuristic algorithms, Support vector machine	Diabetes dataset	Accuracy of Genetic Algorithm (NSGA-II) - 98.2% And accuracy of Multi-Objective particle swarm optimization (MOPSO) is 94.6%, respectively.
12	2019	Cuckoo search, Gabor filters, Mammography, Optimization, Feature extraction	DDSM database	Outperforms some of the best techniques used for mammogram classification based on Sensitivity, Specificity, Accuracy, and Area under the curve

13	2019	Feature selection, Deep learning, Predictive analytics	diabetic, breast cancer and chronic kidney disease datasets	Classification accuracy with minimal time requirement compared to existing Feature selection
14	2019	Cancer, classification, Feature selection, genetic profile, machine learning, reject option	Cancer Dataset	Results shows that for particular dataset the predictive accuracies of RO classifiers were different for all the different Feature Selection methods.
15	2018	Binary Particle Swarm Optimization Feature selection, Healthcare data classification, Swarm intelligence	UCI cancer classification dataset	How much the proposed method is effective on the basis of accuracy & feature selection cost
16	2018	Ensemble Methods, Feature selection, Heart disease, Particle swam optimization	Coronary Heart Disease (CHD)	Results show that Bagged Tree and PSO achieved the highest accuracy
17	2018	Disease prediction, Prediction accuracy, Neural network, Machine learning, Ensemble	Disease prediction related work	Outcomes are checked for accuracy, f-measure, precision and error rate
18	2018	Attribute optimizatio n, Data mining, Dengue, Fitness function, Genetic algorithm	Dengue dataset	Results indicate our algorithm is more efficient and accurate in determining presence of Dengue disease
20	2017	Big Data, Differential evolution, Fuzzy AHP, e- Healthcare, Feature selection, Optimization and Feed-forward neural network	AI, Neural network, AHP related resreach	Accuracy is higher than compared to other models where the accuracy is 83%
21	2017	Coronary disease diagnosis, Multi-objective optimizatio n and	Coronary disease dataset	The given method improves the accuracy of lesion detection for

		Embedded Feature selection		betterment of treatment planning
22	2017	Particle swarm, Medical diagnosis, Evolutionary algorithm, Imbalanced dataset	Imbalanced dataset	M-PSO statistically outperforms the others
23	2016	Lung cancer, Correlation based Feature selection (CFS), Neural network	Lung cancer	Several lung CT images and has proven to attain excellent results in the classification of lung cancers
24	2015	Data Enhancement, Diagnostic Gait, Analysis and Differencing	Gait patterns	Accuracy of average classification is 87.91%
25	2015	Classification, feature selection, normalization	Chronic heart and diabetes	Predicting diseases through medical data mining and healthcare
26	2015	Metaheuristic algorithms, Random tree and Parkinson disease	Parkinson Dataset	In classifying and feature filtering, ABO algorithm shows 97% accuracy
27	2013	Classification, heart sound disease, dimensional reduction	heart dataset	Overall classification accuracy of rough set approach is higher than other ML language.

The outline of this paper is section 1 presents introduction, section 2 details out preliminary data collection, section 3 network analysis, section 4 gives limitations of the carried out study followed by conclusion in section 5. References referred to formulate this paper are at the end.

2. Preliminary Data Collection

Since 2004, Scopus is the world's enormous citation and abstract database of peer-analysis literature, including books, scientific journals and conference papers, comprising of research topics across all scientific and technical direction, medicine and community studies to arts and humanities. In addition, Scopus also provides very smart tools to examine, trace and visualize research to deliver a broad overview of the world's scientific research result across all genres. On the other hand, Web of Science is a platform comprising of many literatures search databases designed to back scientific and scholarly research. It provides a usual search language, navigation domain and data structure. In comparison to Web of science, Scopus has a larger dataset and updates their contents daily whereas the WOS updates weekly.

2.1. Database Search Query

Table 2 The Search query for databases

Scopus DB query	“Feature Selection” AND “Optimization Techniques” AND “healthcare”
Web of Science DB query	

The “Feature selection using optimization techniques in healthcare” query has been used for analysis in database and a comparative study has been performed from Scopus and Web of Science database.

2.2 Initial search results

By employing the query mentioned in Table 1, 53 publications are fetched from the Scopus database. All of the 53 publications are published in the English language. (Table 3).

Table 3 Language of Publications

Publication Language	Publications
English	53

Source: Scopus_DB accessed on 22th November 2020

Most of the publications considered were comprising of articles, conference papers, and conference reviews (Table 4). The highest number of publications have been found in the article publication type, leading with a number of 29 publications, which comprises of 54.7% of the total publication results. Followed by conference paper and conference review with 12 and 12 number of publications respectively with 22.64% of the total publications.

Table 4 Type of Publications

Publication Type	Number of publications	Percentage of 53
Article	29	54.7%
Conference Paper	12	22.64%
Conference Review	12	22.64%
Total	53	100%

Source: Scopus_DB accessed on 22th November 2020

The 53 publications fall under various source types such as Journal, Book Series, and Conference proceeding. 50.94% of the source type comprises the Journal source type, leading with a number of 27 sources. Book series is followed by a percentage of 30.18% with a number of 16 sources. 10 number of sources of the total results comprise 18.86% fall under the conference proceeding source type, which is the lowest of the three source types.

Table 5 Types of Sources for Publications

Source Type	Number of Sources	Percentage of 53
Journal	27	50.94%
Book Series	16	30.18%
Conference Proceeding	10	18.86%
Total	**53**	**100%**

Source: Scopus_DB accessed on 22th November 2020

2.3 Document Types

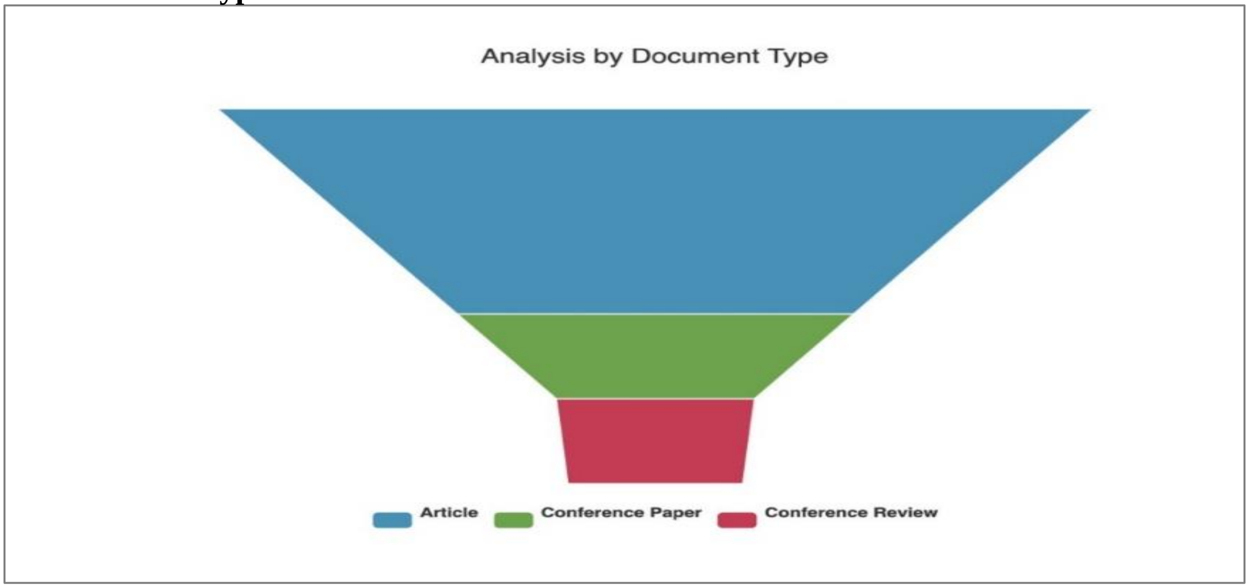

Fig. 1 Types of Document for Publications

Source: Scopus_DB accessed on 22th November 2020

The analysis was done based on different document types. The fig. 1 shows the breakdown of document types from the Scopus database, and fig. 2 shows the classification of document types done by using the Web of Science database. The comparison from both charts shows that the maximum documents are of article type and the least number of records are of conference review type.

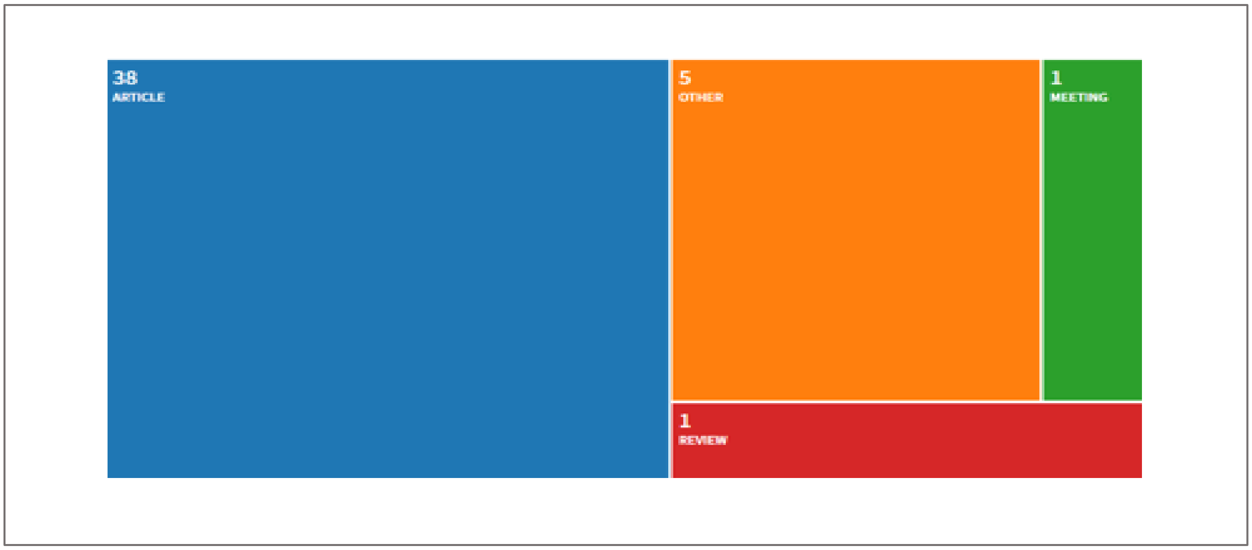

Fig. 2 Types of Document for Publications

Source: Web of Science DB accessed on 22th November 2020

2.4 Preliminary Data Collection

The Charts show the yearly analysis of Scopus (Fig. 3) and Web of Science (Fig. 4) publications. Both the charts show that the combined publications from the year 2020 and 2019 contribute to the highest number of publications, nearly equal to the publications combined from the year 2003 to 2018.

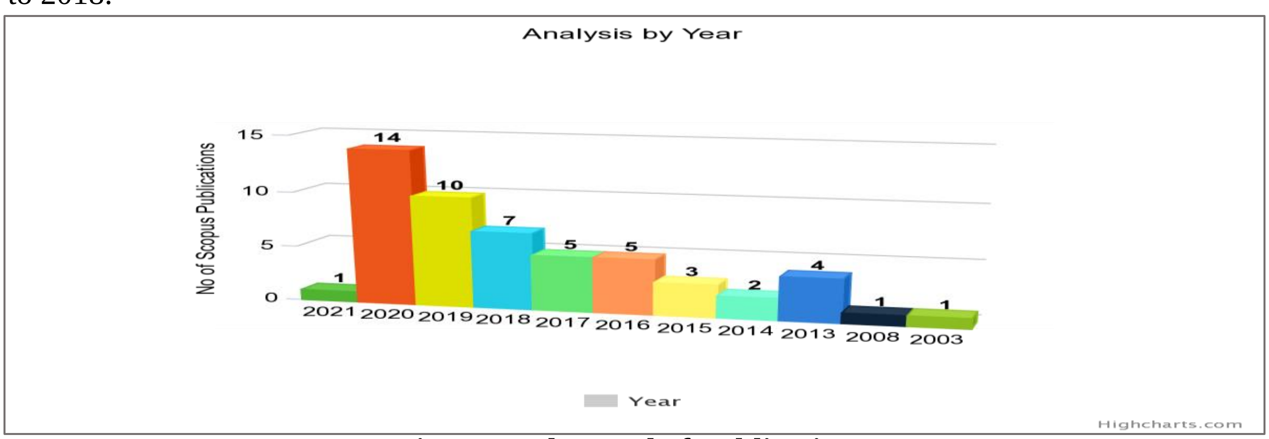

Fig. 3 Yearly trend of Publications

Source: Scopus_DB accessed on 22th November 2020

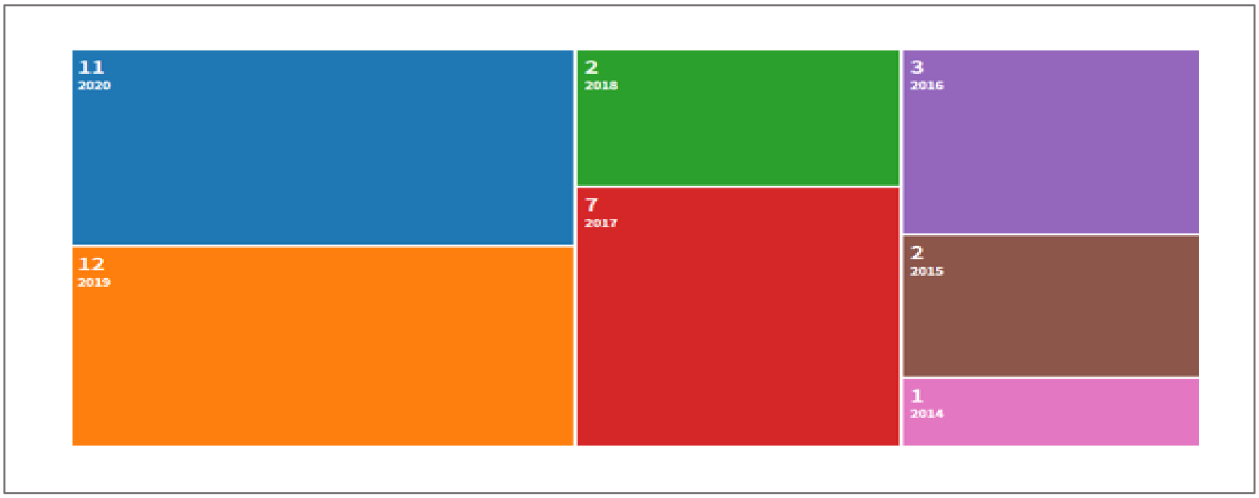

Fig. 4 Yearly trend of Publications

Source: Web of Science DB accessed on 22th November 2020

The charts show the number of publications done by different countries. Notably, India heads all other countries by publishing the maximum number of research documents in Scopus's analysis (Fig. 5) and Web of Science analysis (Fig. 6).

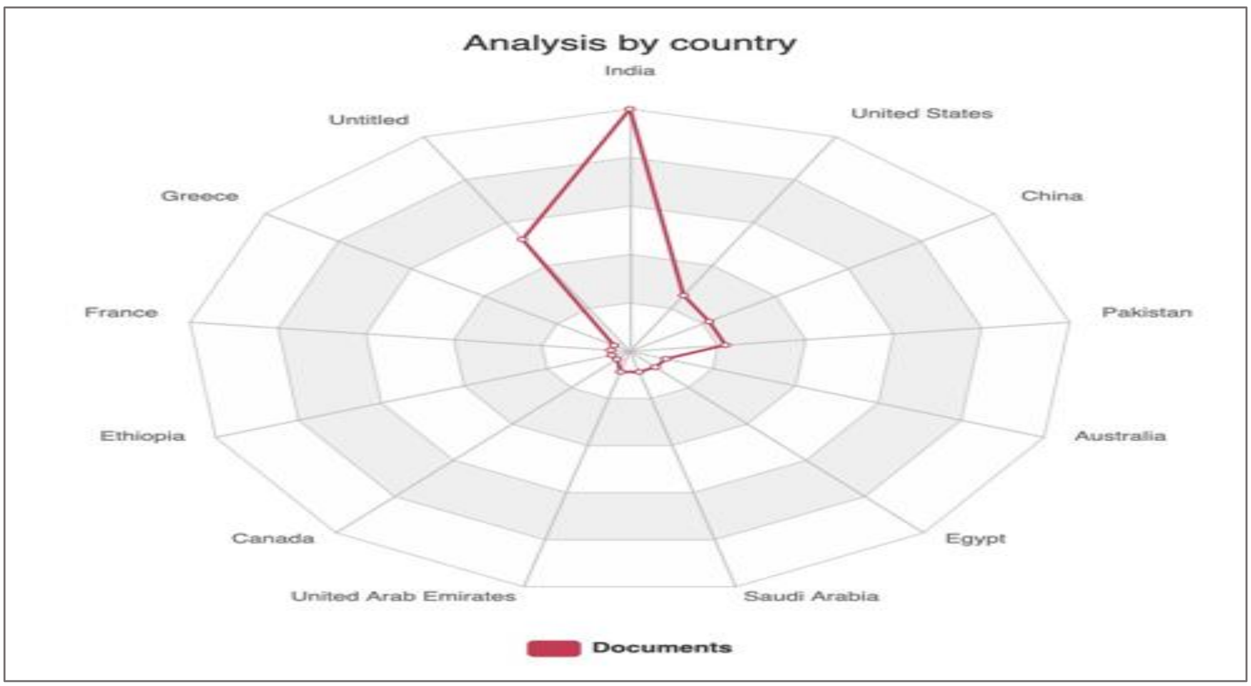

Fig. 5 Countries contributing to the Publications

Source: Scopus_DB accessed on 22th November 2020

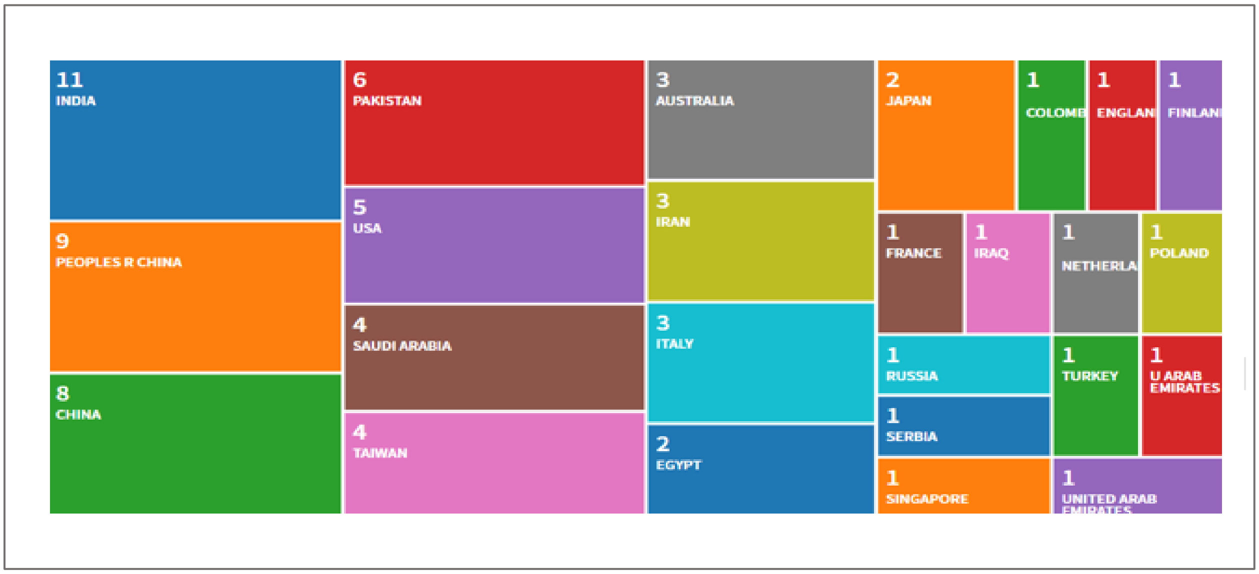

Fig. 6 Countries contributing to the Publications

Source: Web of Science DB accessed on 22th November 2020

Documents related to feature selection using optimization techniques in healthcare have been collected from 2008 to 2021 for 13 years. The below table 6 exhibits trends in the number of publishing counts per year in the feature selection using optimization techniques in the healthcare research area. By analyzing this data, the number of publications has been higher in the year 2020, followed by 2019 and 2018.

Table 6 Yearly Publications

Year	Publication Count	Year	Publication Count
2021	1	2016	5
2020	14	2015	3
2019	10	2014	2
2018	7	2013	4
2017	5	2008	1
Total		**53**	

Source: Scopus_DB accessed on 22th November 2020

Advances in Intelligent Systems and Computing are the leading source with a total of 11 publications in Scopus DB (Figure 4.1), whereas the prominent source in Web of Science DB (Figure 4.1) is IEEE Access with ten publications.

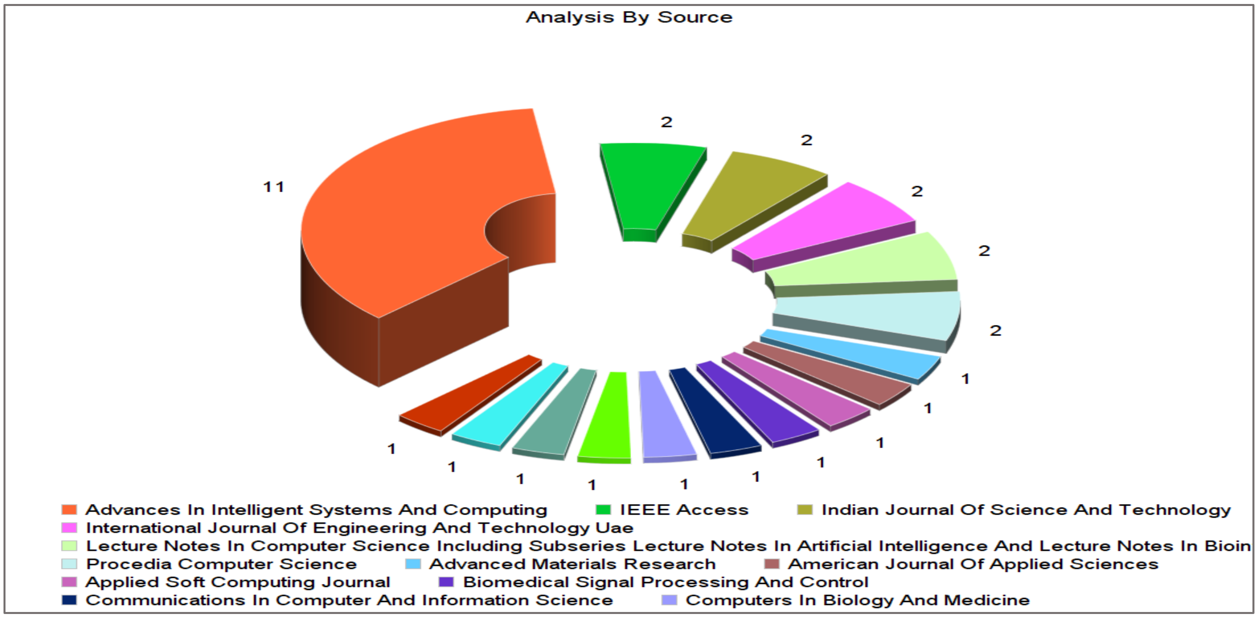

Fig. 7 Publication statistics of source

Source: Scopus_DB accessed on 22th November 2020

Fig. 8 Publication statistics of source

Source: Web of Science DB accessed on 22th November 2020

2.5 Keyword Statistics

Researchers scheme the usage of appropriate keywords to search within the database. The placement of exact keywords is required to get the crucial research areas. The below table shows the top ten keywords list, which are considered from publications in feature selection. It gives us an overview of the number of publications based on significant keywords. Feature extraction, feature selection are more attentive keywords and Diseases, PSO are less attentive keywords.

Table 7 Top ten keywords for feature selection

Keywords	Number of Publications
Feature Extraction	19
Feature Selection	16
Data Mining	12
Diagnosis	12
Health Care	12
Machine Learning	11
Classification	8
Optimization	8
Diseases	7
Particle Swarm Optimization (PSO)	7

Source: Scopus_DB accessed on 22th November 2020

3. Network Analysis

3.1. Network Diagrams

A network diagram stands in for the graphical and visual representation of the connection between the nodes, based on different parameters. The network analysis of Figure 9 to 14 is carried out using various tools such as Gephi, Minivans, Sciencescape, and VOSViewer. Most of the network analysis is performed using Gephi and Minivans. The analysis diagrams from both the tools are shown to better understand the connections and depict a comparative analysis of both the software.

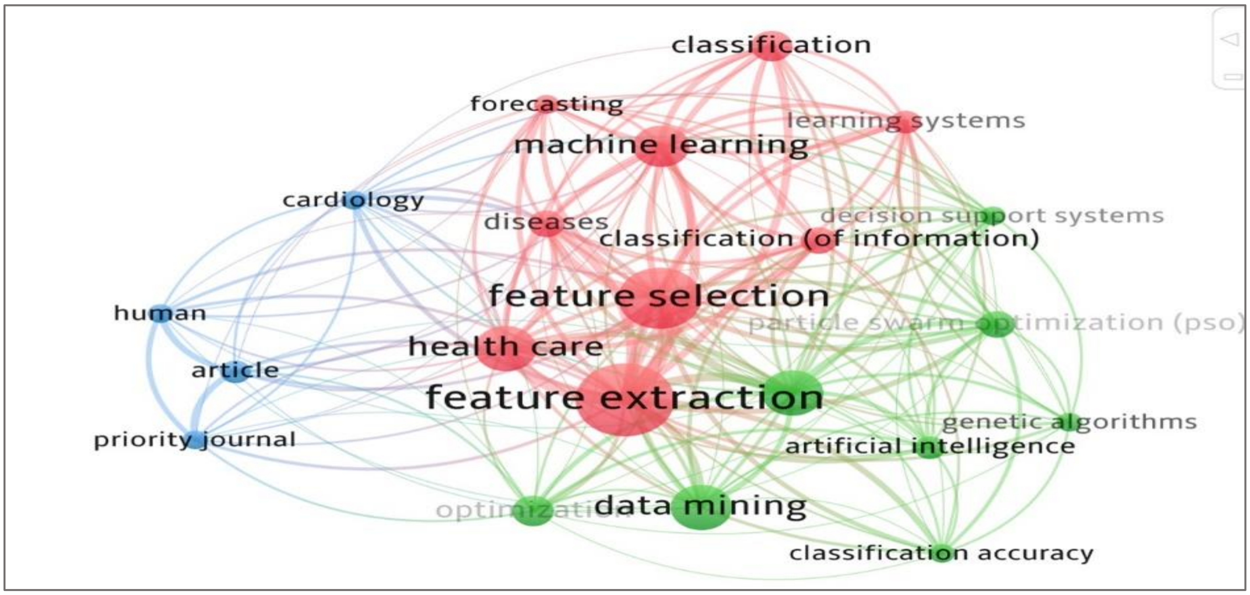

Fig. 9 Cluster of Co-appearing Index Keywords

Fig. 9 shows the cluster of co-appearing index keywords, which consists of 21 items forming three different sets of groups represented by different colors, and the network has connected through 173 links. From the cluster, it is very evident that 'feature selection' and' feature extraction, are significant keywords in the network.

The fig. 10 represents a network of papers that linked by citation when they have a DOI. The network does not connect through edges, the differentiation of the 36 nodes has been done based on each document's title, and each paper has contributed an equal percentage to the network.

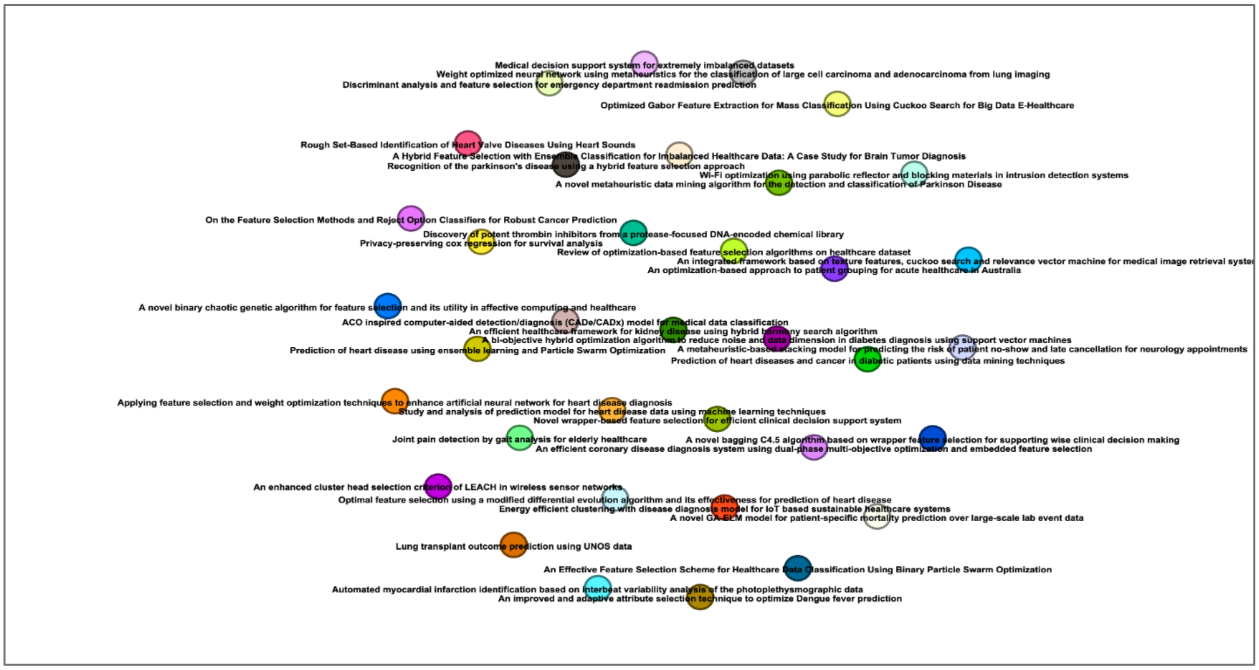

Fig. 10 Papers linked by citation

The fig. 11 and 12 are the comparative network diagrams of reference scape from the minivan and Gephi software. In the cluster, references cited in less than two papers are filtered out. References connected to less than 0 other references are filtered out. Authors present in less than three articles are filtered out. Also, Keywords present in less than three papers are filtered out. Additionally, Journals present in less than five articles are filtered out, and lastly, Networks are merged, and nodes with less than two neighbors are filtered out recursively. After the filtering, the structure ends up with 45 nodes and 235 edges.

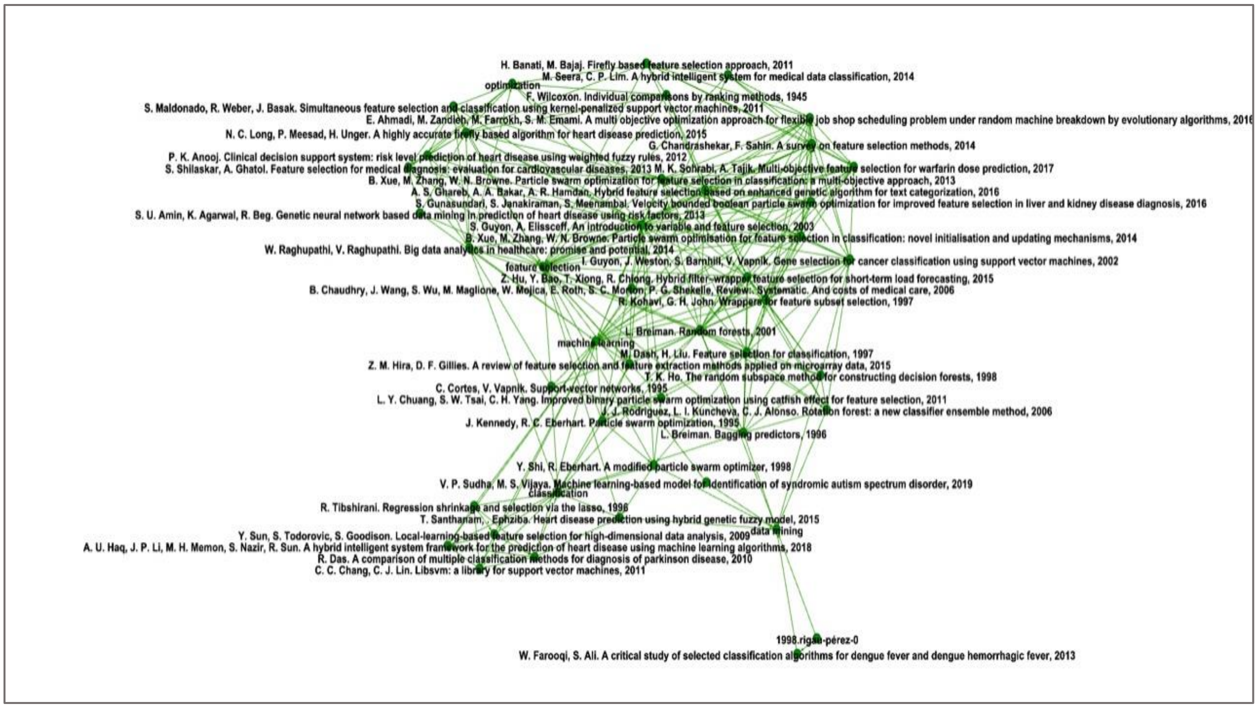

Fig. 11 Reference Scape (Drawn using Gephi)

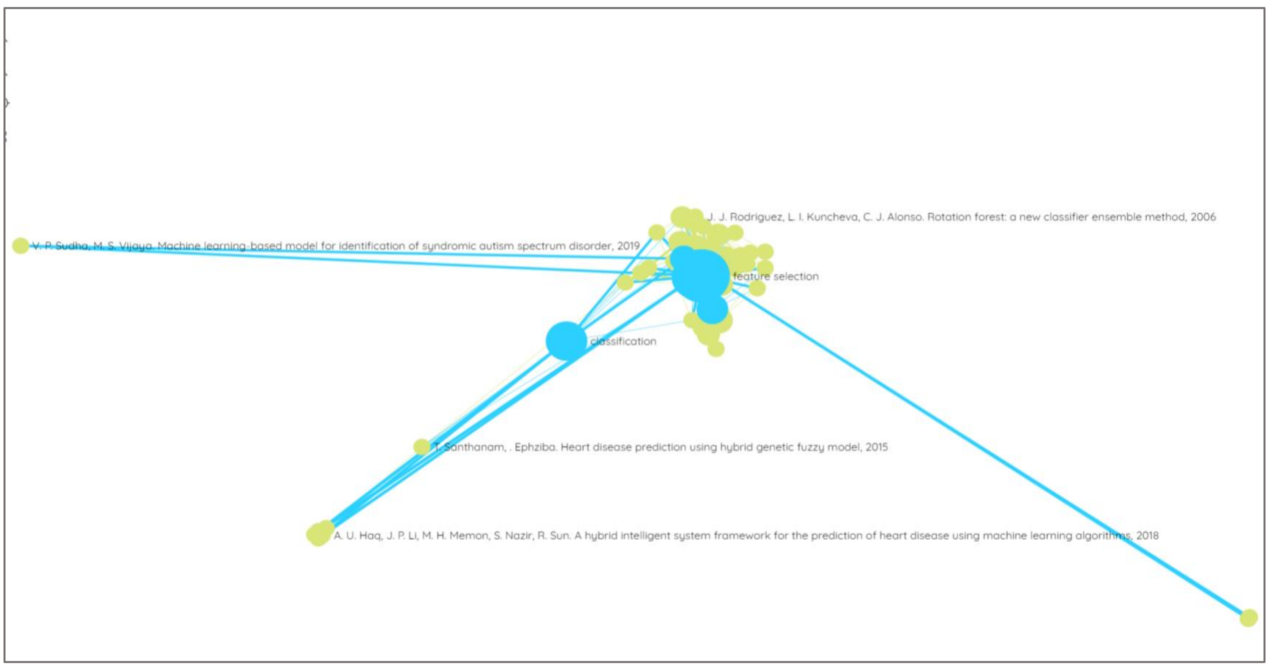

Fig. 12 Reference Scape (Drawn using Minivan)

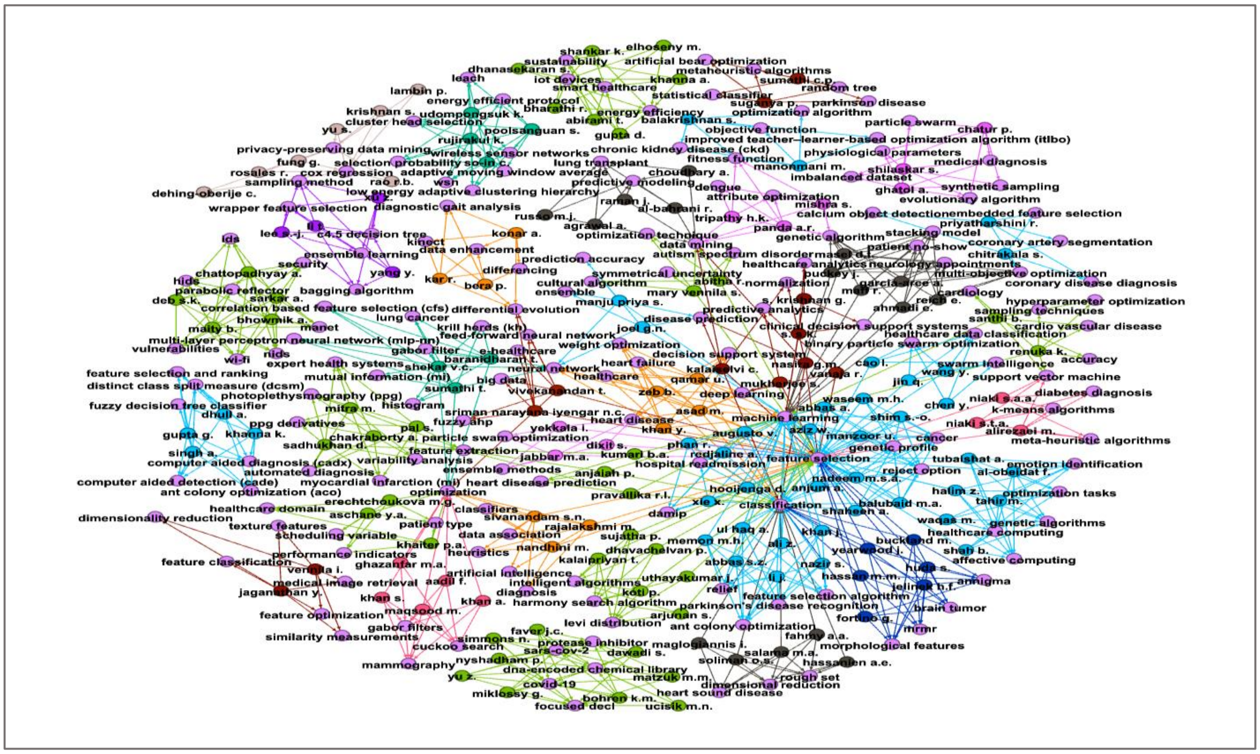

Fig. 13 Author and Author Keyword (Co-appearing in the same paper)

Fig. 13 shows the linkage between author and author keywords that are co-appearing in the same paper with 859 edges which connect 331 nodes. The maximum number of citations received by a publication is 45. Fig. 14 and Fig. 15 shows a comparison of the network diagrams drawn using Gephi and Minivan, respectively. For filtering, three nodes are removed with remove disconnected nodes mode.

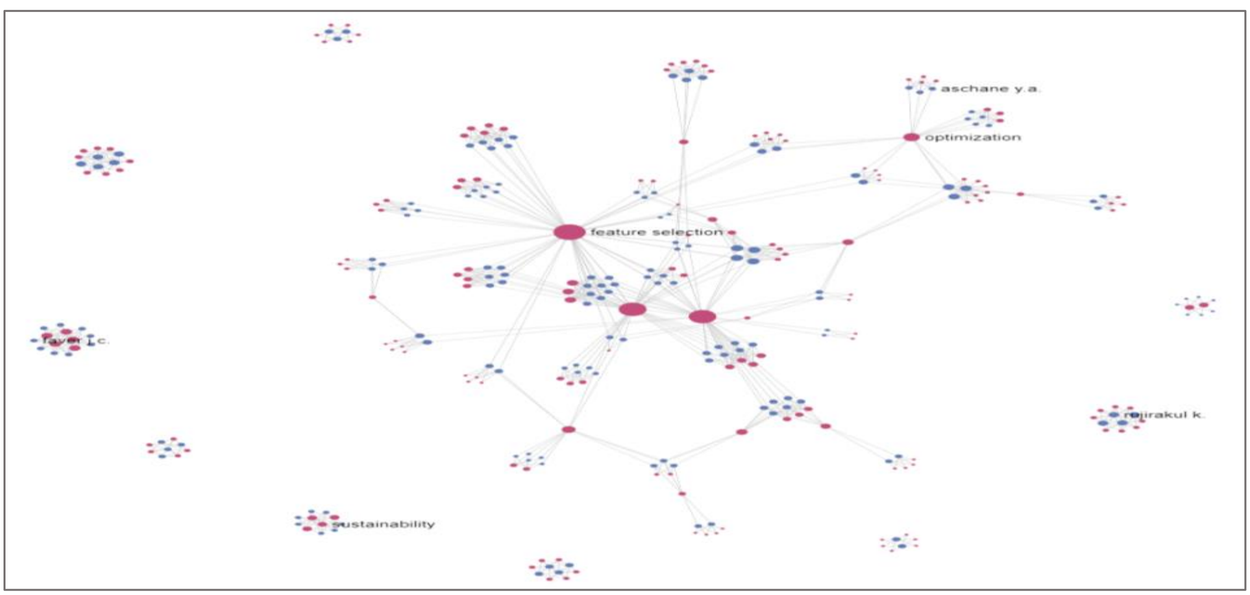

Fig. 14 Author and Author Keywords, Co-appearing in the same paper

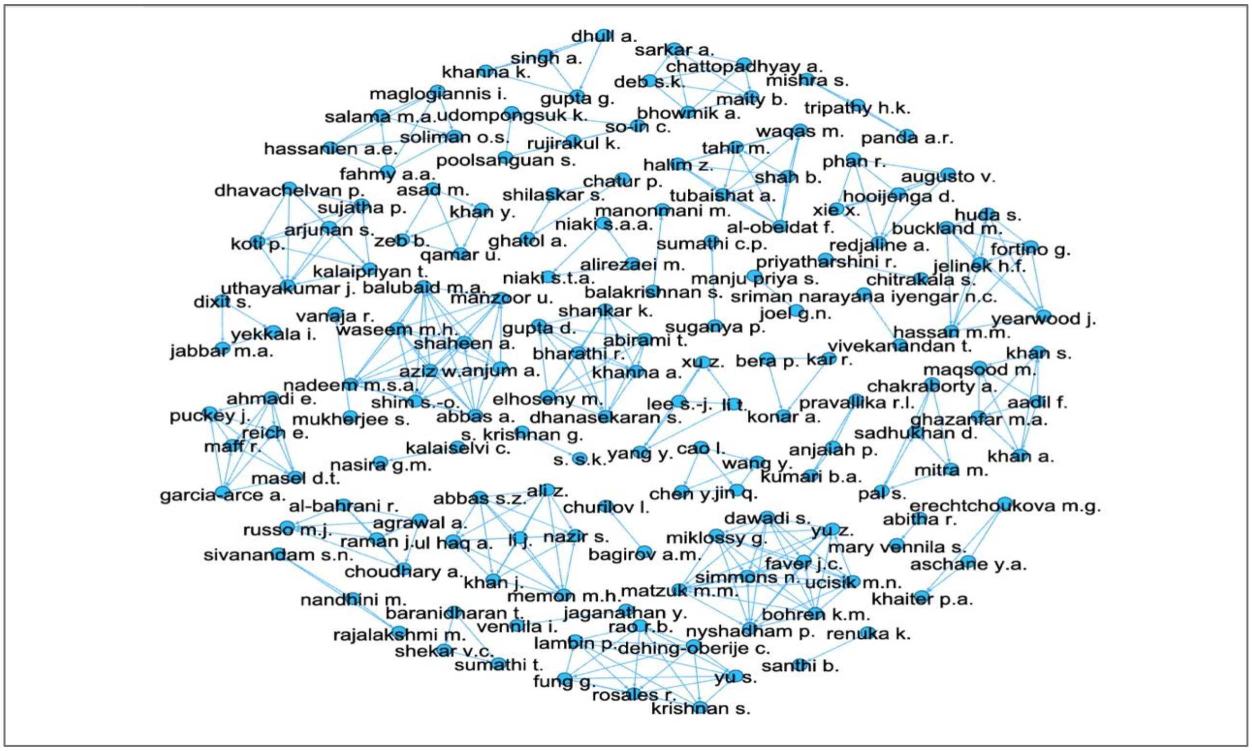

Fig. 15 Network of authors linked by co-publication

The relation between authors and their co-published papers is depicted through Fig. 16 and Fig. 17 network diagrams. Fig. 17 throws light on the prominent authors with co-publications, which is made using minivan, whereas Fig. 18 made with Gephi clearly depicts the networks between these authors linked by co-publications. The layout shows 163 nodes and 320 edges with one disconnected node removed as a filter.

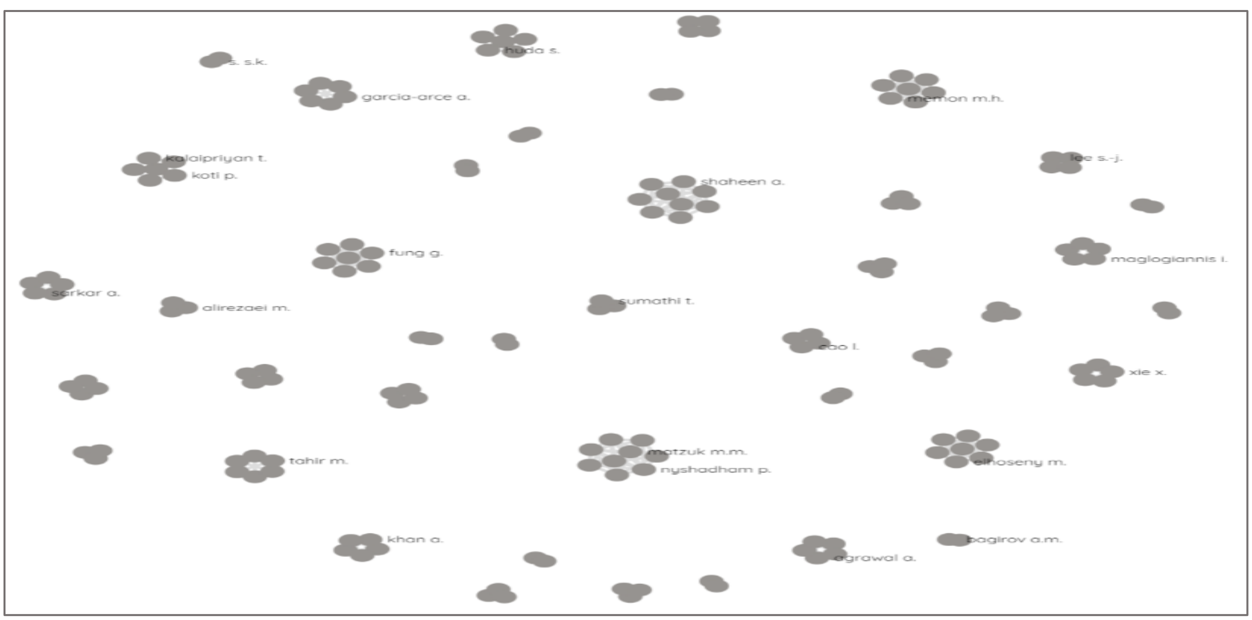

Fig. 17 Network of authors linked by co-publication

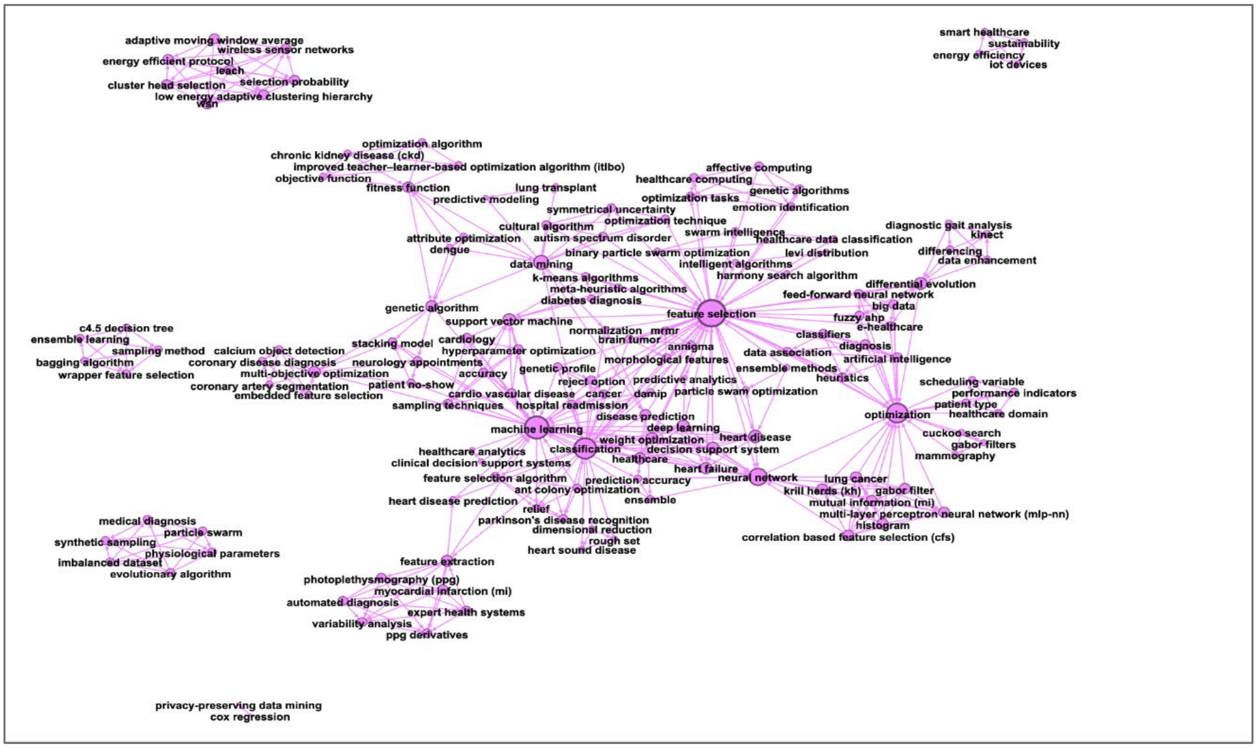

Fig. 18 Author keyword, co-appearing in the same paper

Fig. 19 and 20 depicts a network connection between author keywords that are co-appearing in the same papers in which, the represents 170 nodes connected with 513 edges. “Feature selection”, “machine learning” and “classification” are the primary keywords used in this area. Fig. 19 has been made using Gephi software and shows a clustered representation of the author keywords co-appearing in the same paper. Fig. 20 has been made using Minivan software and shows a clustered spread out representation of the author keywords that are present in the same paper.

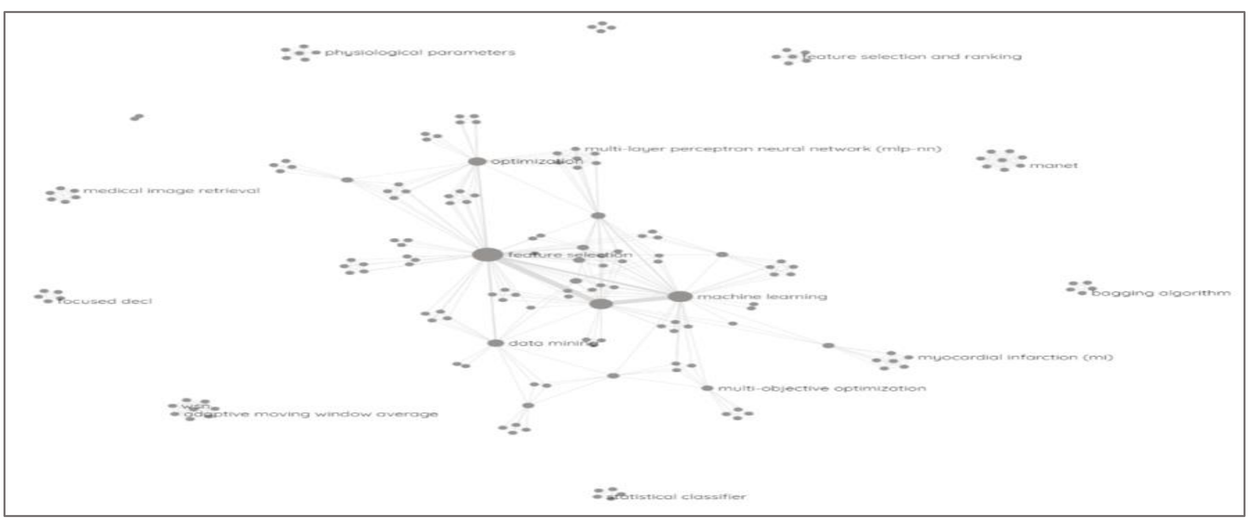

Fig. 19 Author keyword, co-appearing in the same paper

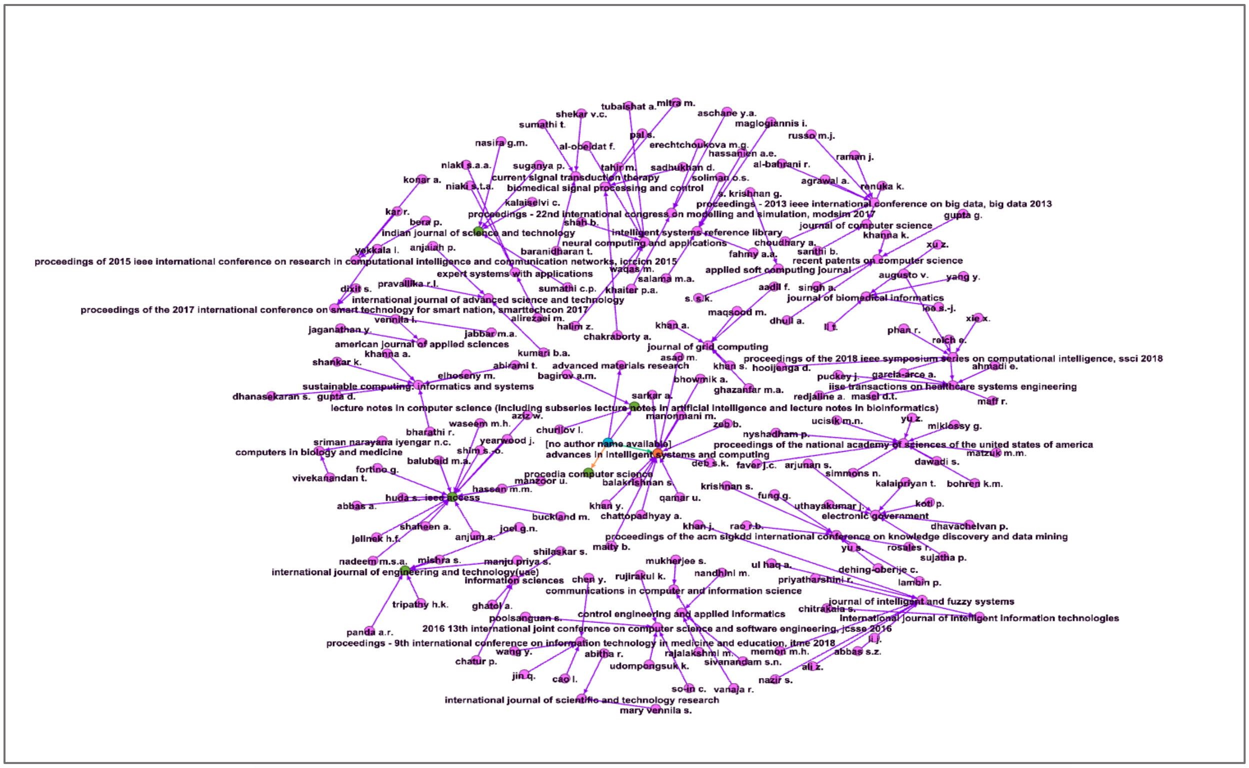

Fig. 20 **Authors and Source Titles, co-appearing in the same papers**

Fig. 21 and 22 depicts clustering and connection of author keywords and source titles that are co-appearing in the same papers with 202 nodes and 167 edges. The different colours in the diagram represent the different number of papers written by individual authors.

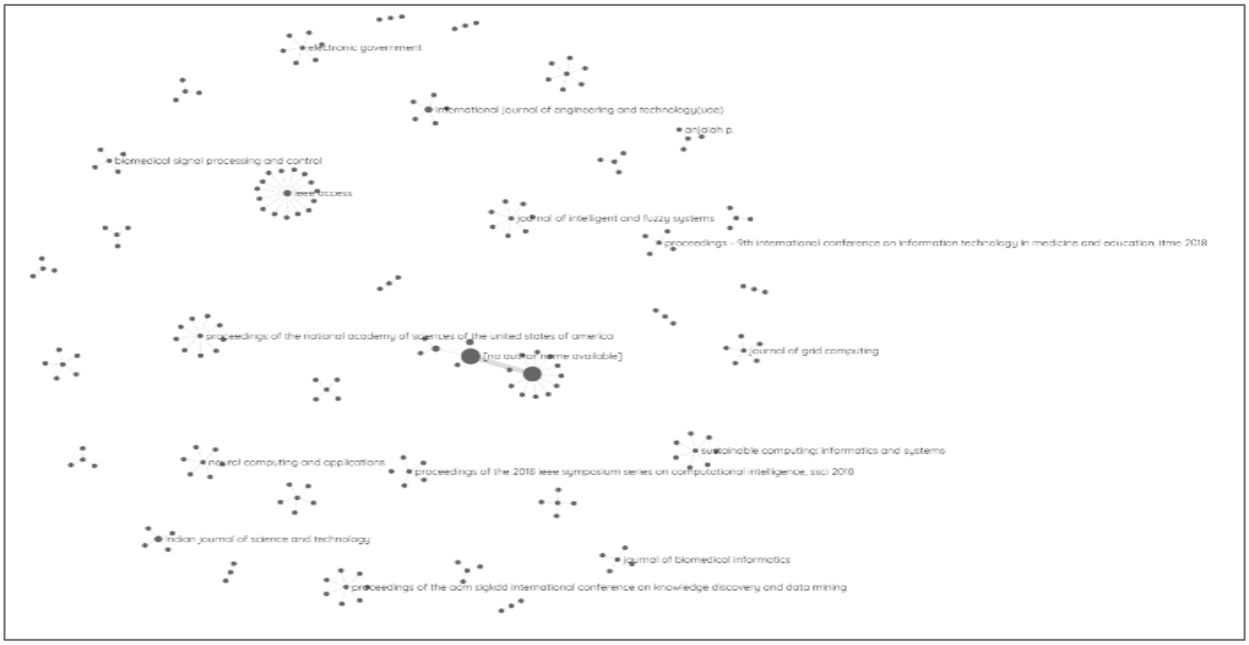

Fig. 21 **Authors and Source Titles, co-appearing in the same papers**

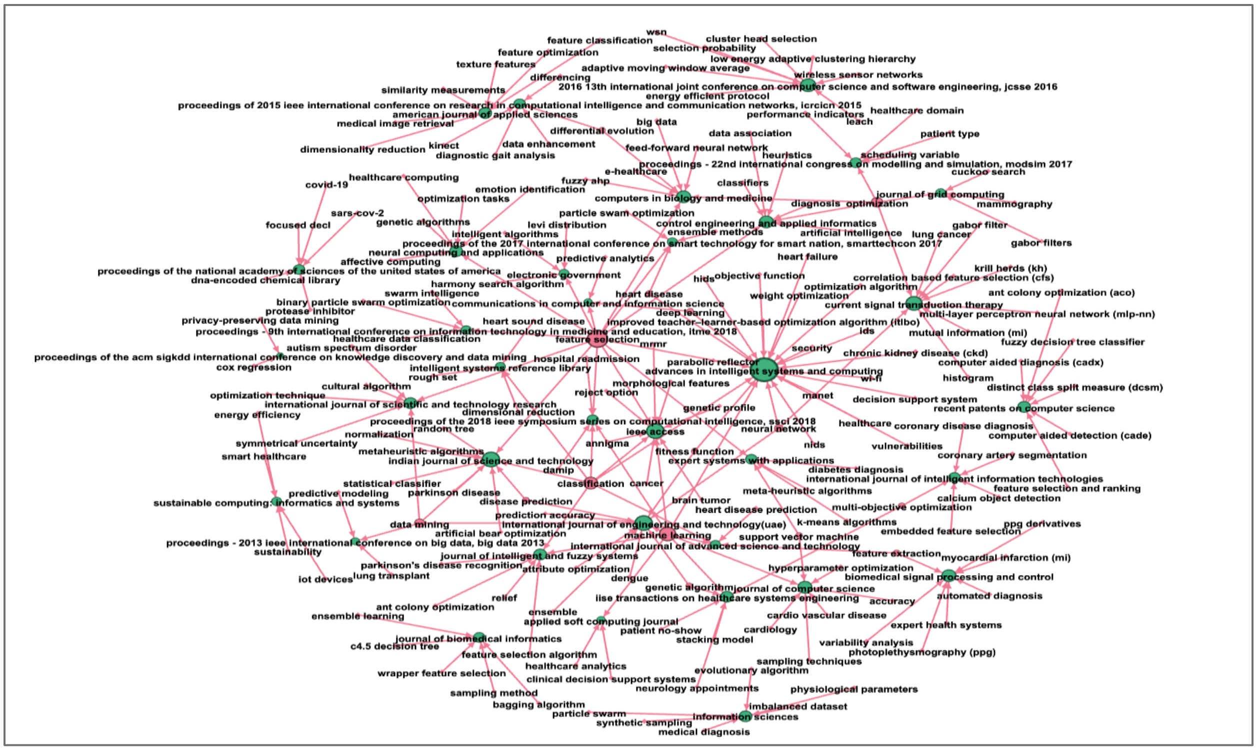

Fig. 22 Co-appearance of author keywords and source title in the same paper

Fig. 21 shows a clustered diagram showing titles of the source and author keywords, that are co-appearing in the same publications with 214 edges connecting 205 nodes and made in Gephi. The pink nodes represent the author keywords and green nodes represents source title, and edges form the link between them. In the filtering of the disconnected nodes, three nodes have been removed. Fig. 22 represents the same cluster in Minivan.

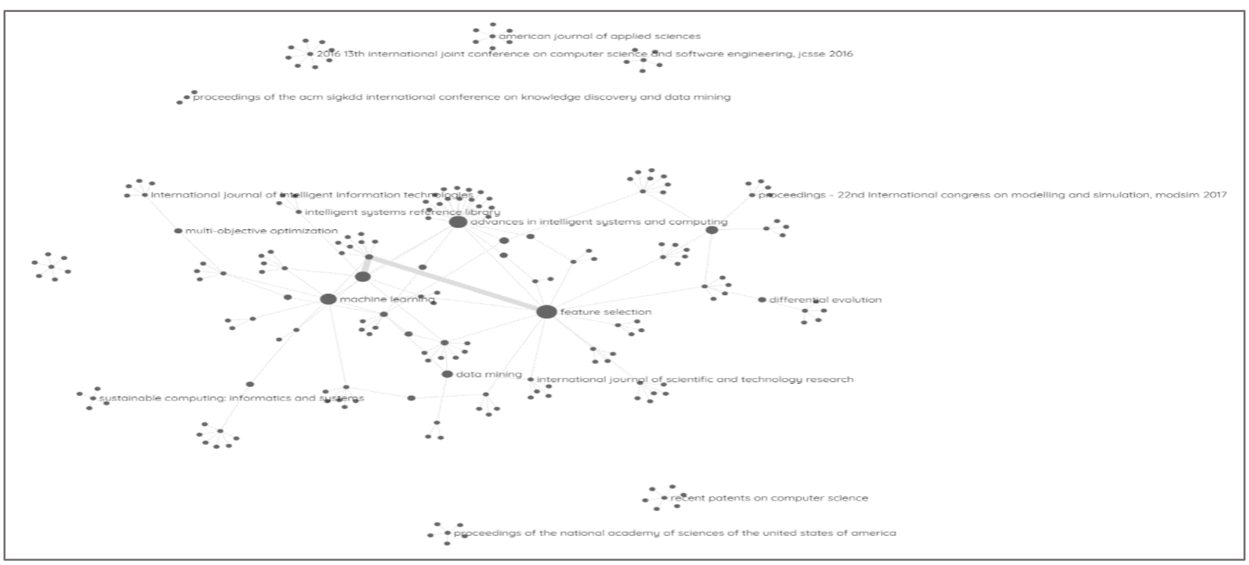

Fig. 23 Co-appearance of author keywords and source title in the same paper

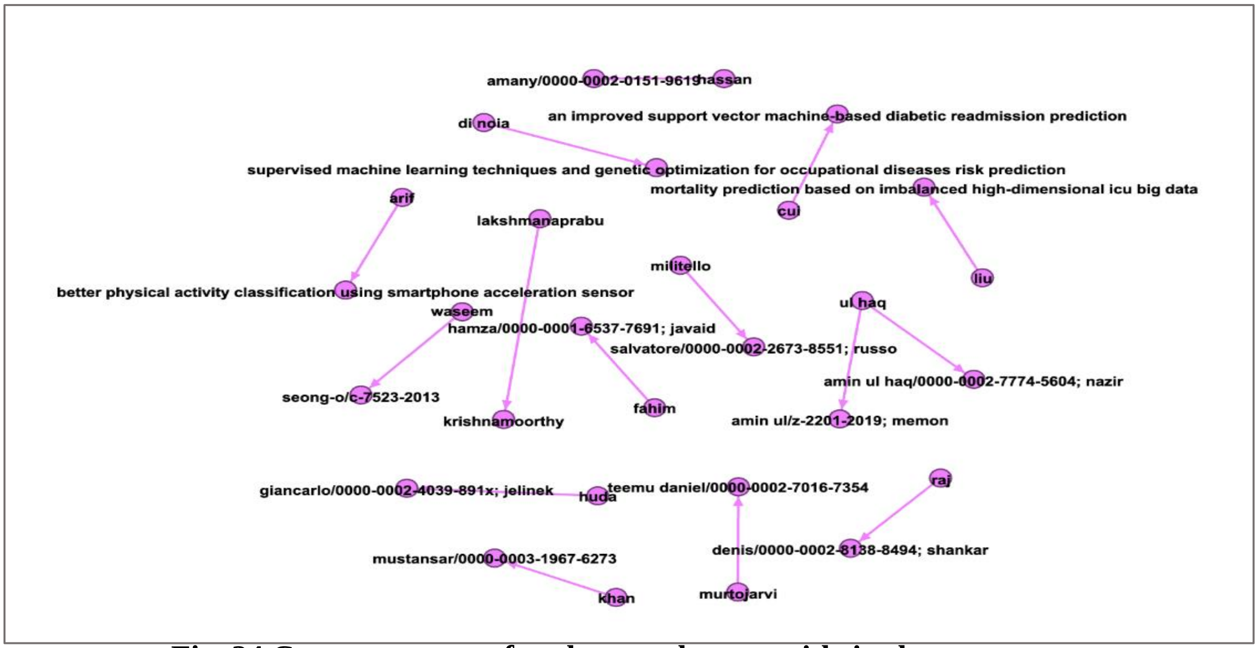

Fig. 24 Co-appearance of authors and source title in the same paper

Fig. 23, 24 show a clustered network analysis of the authors and source title that are co-appearing in the same publications, based on the publications in Web of Science database. The fig. shows 29 nodes and 15 edges. After using the filter to remove disconnected nodes, 22 nodes were removed.

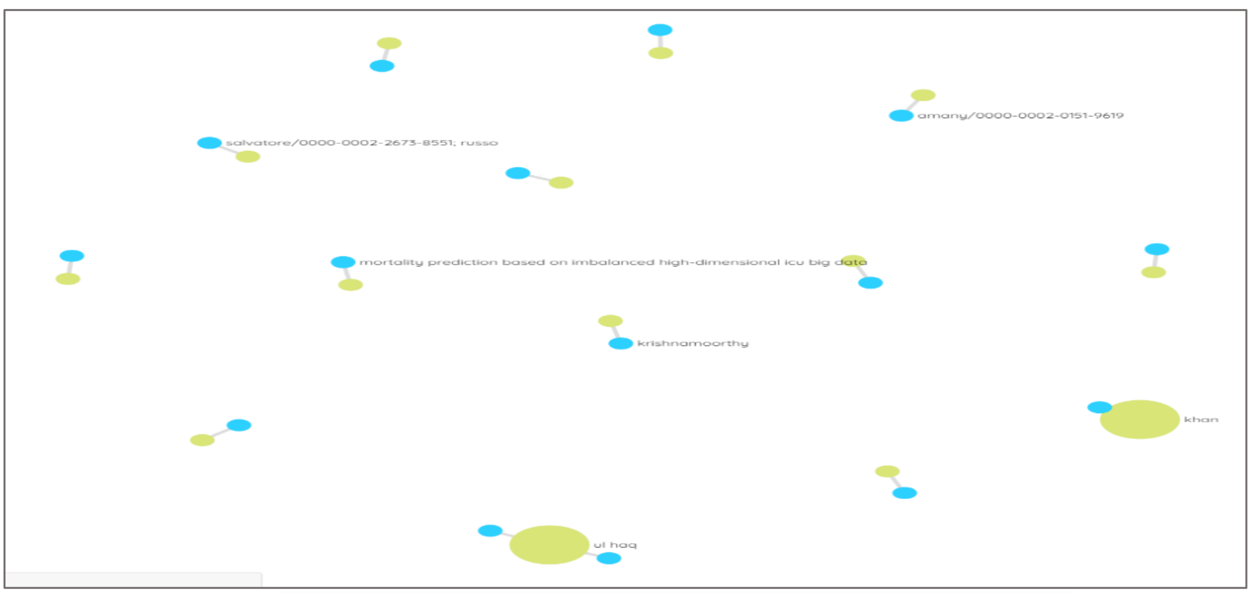

***Fig. 23** Co-appearance of authors and source title in the same paper*

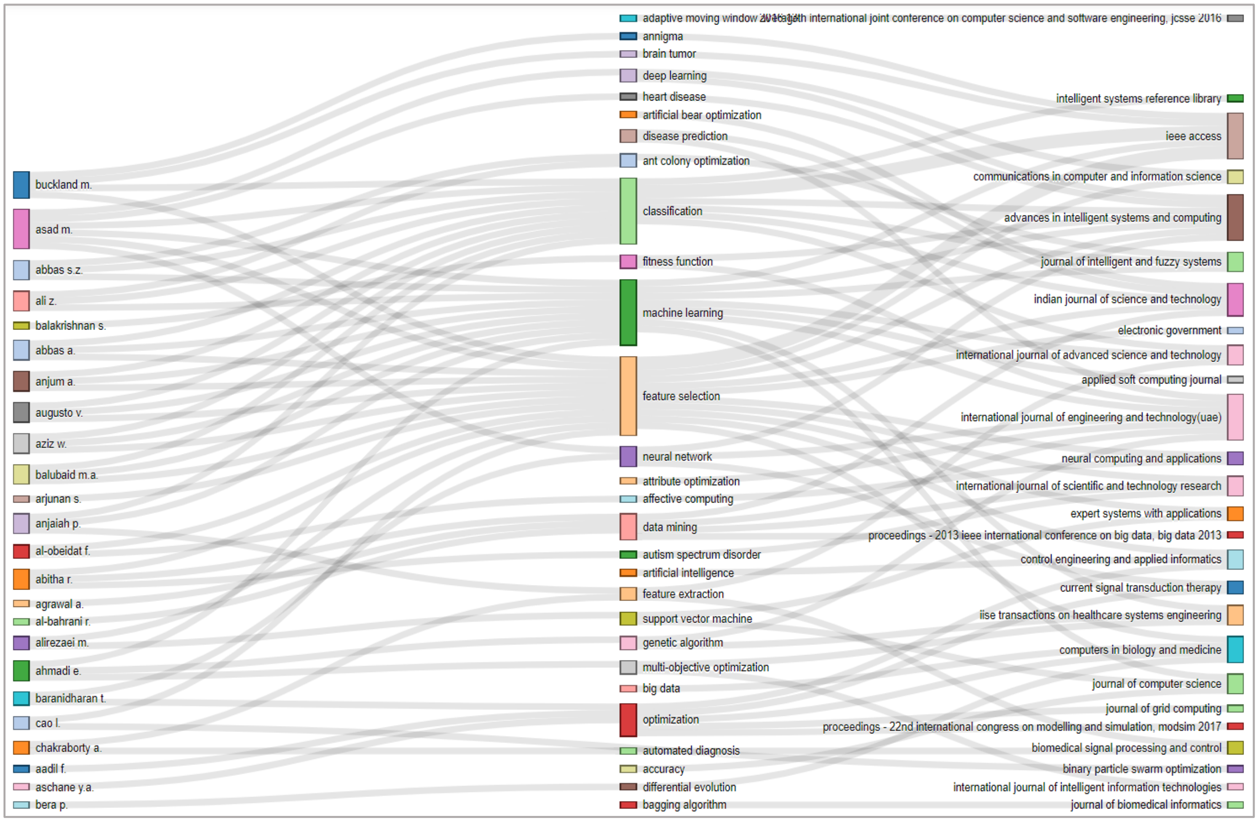

Fig. 24 Connection between Network of Main-authors, Keywords, and Journals

Fig. 24 gives the general overview of the main authors, primary keywords, and prominent journals from "Feature selection using optimized techniques healthcare".

3.2 Affiliation Statistics

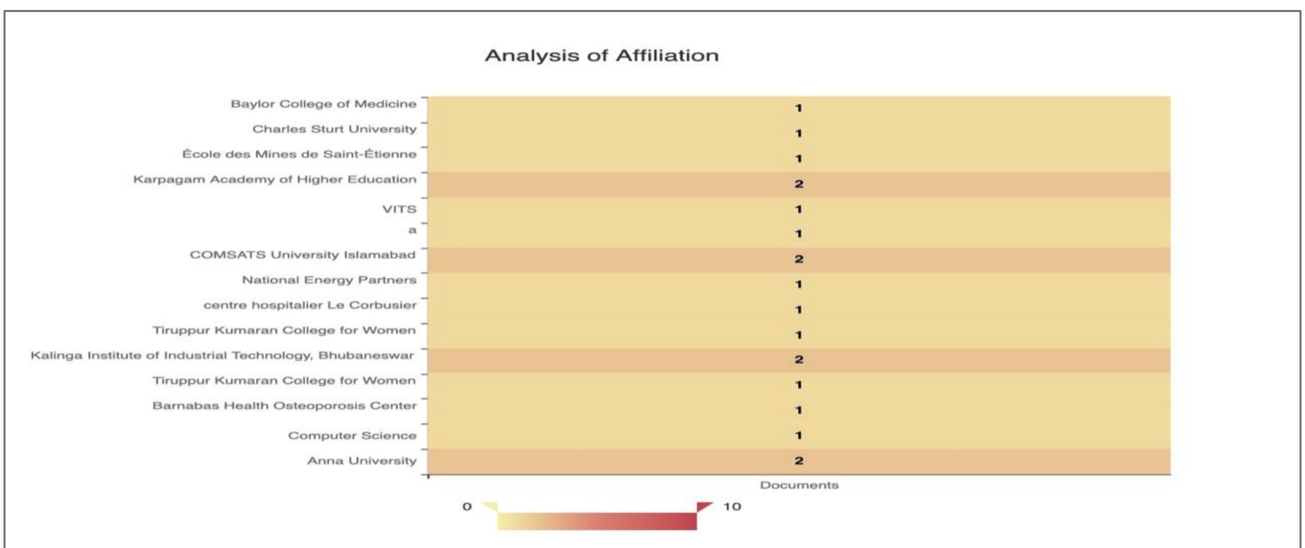

Fig. 25 Analysis by Affiliation

Source: Scopus_DB accessed on 22th November 2020

University/ Organization affiliation statistics for the publications in Scopus database (Fig. 25) indicate that three of them are situated in India among the top universities. The publications obtained from Web of Science (Fig. 26) the top two universities/ organizations are COMSATS University, Islamabad, and Deakin University, Australia.

Fig. 26 Analysis by Affiliation

Source: Web of Science DB accessed on 22th November 2020

3.3 Citation Analysis

Table 8 Analysis for citations of publications

Year	<2016	2016	2017	2018	2019	2020	>2020	**Total**
Number of Citations	17	16	17	48	73	122	6	**299**

Source: Scopus_DB accessed on 22th November 2020

Table 4 shows the yearly citations obtained from publications extracted in the area of feature selection and optimized technique, the total number of citation count of 53 publications in 299 to date. Evidently, 2020 leads by having the highest number of citations, followed by the year 2019.

3.4 Geographical Regional Analysis

Fig. 27 is drawn using excel charts showing geographic locations of the published papers. The black pointers in the map point to the countries with publications in Scopus. It is evident from the map and the publication statics divided with respect to country/ territory mentioned above. Most numbers of publications are from India, Asia, with 43.3% of publications, followed by the United States with 11% of publications.

Fig. 27 Geographical Locations

3.5 Subject Areas

Fig. 28 and Fig. 29 show the severance based on the subject area from the data extracted from Scopus and web of science datasets, respectively. It is clear from the comparison between both graphs that maximum research is carried out in the computer science area. However, the number of publications on different subject areas slightly differs when compared between both the graphs.

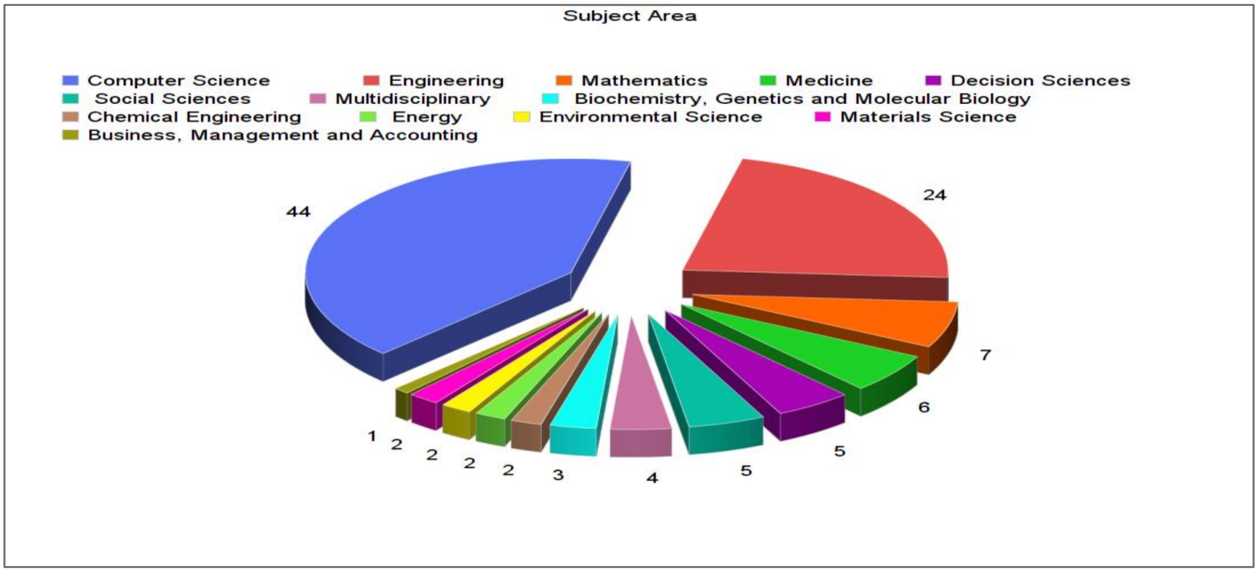

Fig. 28 Subject Areas of publications

Source: Scopus_DB accessed on 22th November 2020

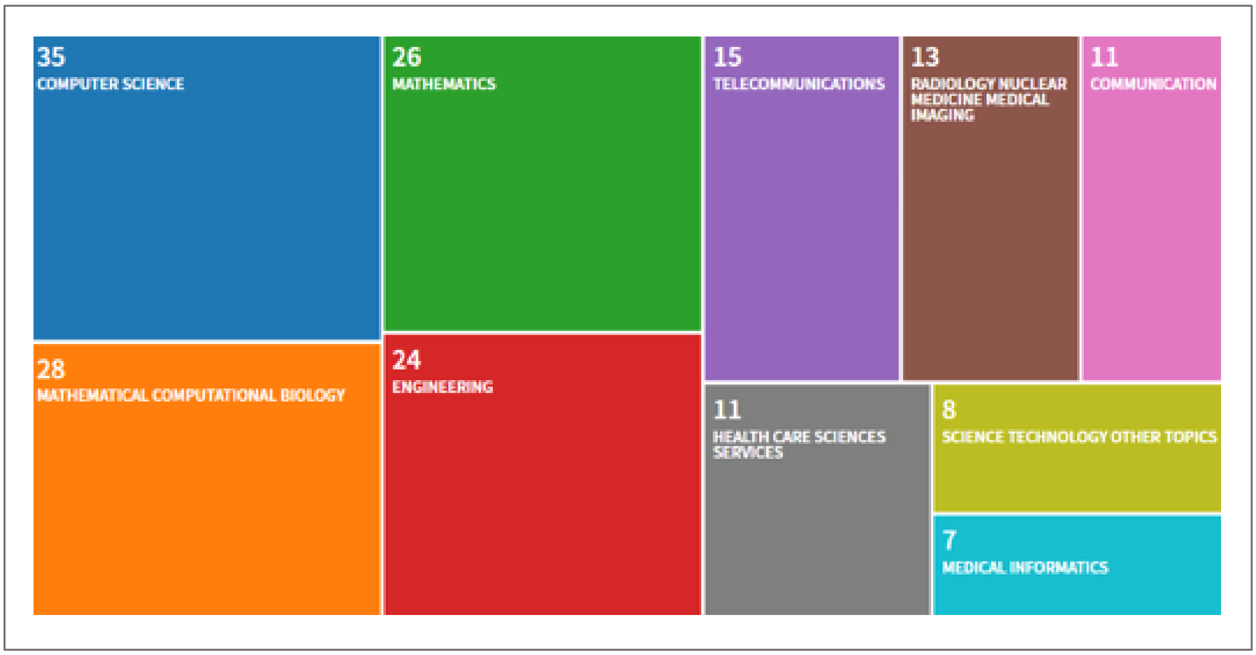

Fig. 29 Subject Areas of publications

Source: Web of Science DB accessed on 22th November 2020

3.6 Analysis by Author

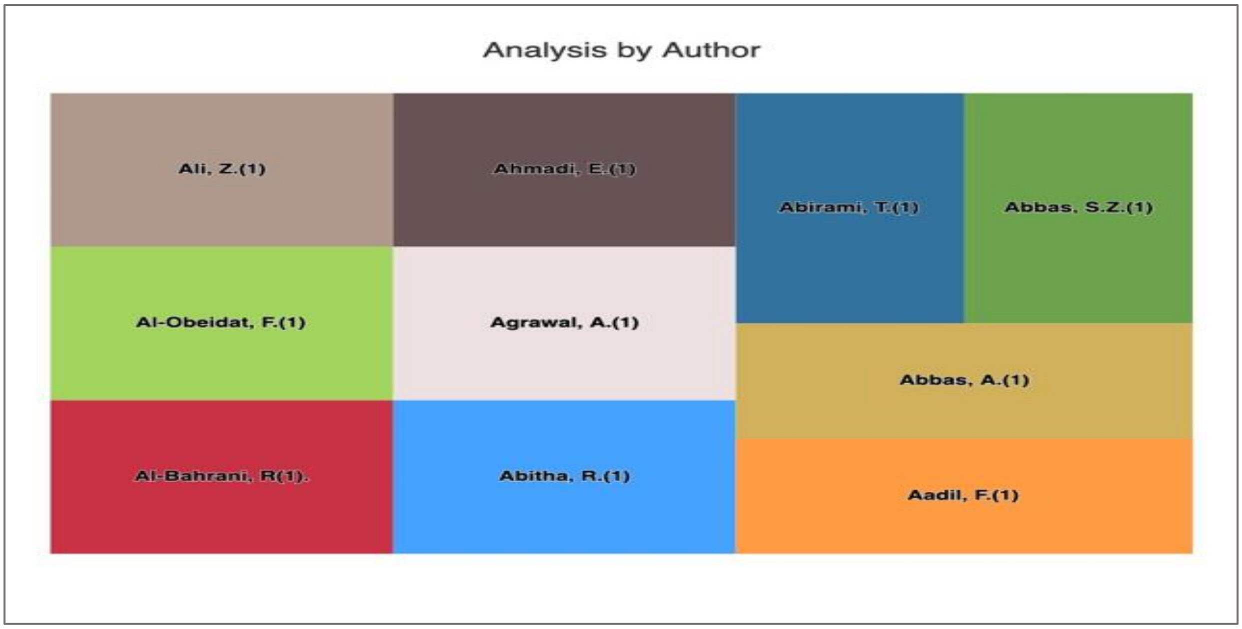

Fig. 30 Main contributing Authors

Source: Scopus_DB accessed on 22th November 2020

Fig. 30 represents the top 10 Authors obtained from the publications in Scopus DB, whereas Fig. 31 depicts authors obtained from the publications in Web of Science DB, the leading author for the publications in feature selection using optimized techniques is Shankar K.

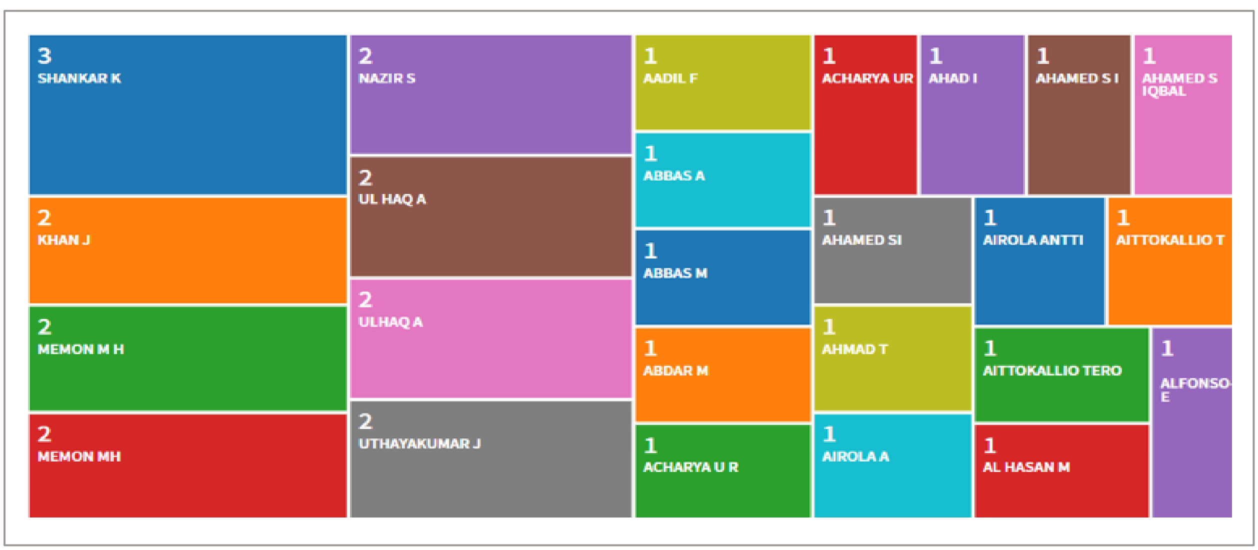

Fig. 31 Main contributing Authors

Source: Web of Science DB accessed on 22th November 2020

3.7 Source Title Citation Analysis

The Table 9 shows yearly source titles, which are obtained from the publications mined in the area of feature selection using optimization techniques in healthcare. The total citation count is 54 to date. The table enumerates the papers, along with the year and total citations.

Table 9: Citation Analysis for Source Titles for publications in Scopus DB

Source Title	<=2015	2016	2017	2018	2019	2020	>2020	Total Citations
Advance in intelligent systems and computing	2	0	0	1	3	5	1	11
IEEE Access	0	1	0	0	1	0	0	2
Indian Journal of Science and Technology	2	0	0	0	0	0	0	2
International Journal of Engineering and Technology Uses	0	0	0	2	0	0	0	2
Lecture notes in Computer Science including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics	1	1	0	0	0	0	0	2
Procedia Computer Science	1	0	0	1	0	0	0	2
13th International Joint Conference on Computer Science and software Engineering JCSSE 2016	0	1	0	0	0	0	0	1
Advanced Materials Research	1	0	0	0	0	0	0	1
American Journal of Applied Sciences	1	0	0	0	0	0	0	1
Applied Soft Computing Journal	0	0	0	0	1	0	0	1
Biomedical Signal Processing And Control	0	0	0	0	0	1	0	1
Communications In Computer And Information Science	0	0	0	0	1	0	0	1
Computers In Biology And Medicine	0	0	1	0	0	0	0	1
Control Engineering And Applied Informatics	0	0	1	0	0	0	0	1

Current Signal Transduction Therapy	0	1	0	0	0	0	0	1
Electronic Government	0	0	0	0	0	1	0	1
Expert Systems With Applications	0	0	0	0	1	0	0	1
Health And Technology	0	0	0	0	0	1	0	1
IISE Transactions On Healthcare Systems Engineering	0	0	0	0	1	0	0	1
Information Sciences	0	0	1	0	0	0	0	1
Intelligent Systems Reference Library	1	0	0	0	0	0	0	1
International Journal of Advanced Science and Technology	0	0	0	0	0	1	0	1
International Journal of Intelligent Information Technologies	0	0	1	0	0	0	0	1
International Journal of Science and Technology	0	0	0	0	0	1	0	1
Journal of Biomedical Informatics	0	0	0	1	0	0	0	1
Journal of Computer Science	0	0	0	0	0	1	0	1
Journal of Grid Computing	0	0	0	0	1	0	0	1
Journal of Intelligent And Fuzzy Systems	0	0	0	0	0	1	0	1
Neutral Computing and Applications	0	0	0	0	0	1	0	1
Proceedings 2013 IEEE International Conference on big data 2013	1	0	0	0	0	0	0	1
Proceedings 22nd International Congress on Modelling and Simulation 2017	0	0	1	0	0	0	0	1
Proceedings 9th International Conference on Informational Technology in Medicine and Education ITME 2018	0	0	0	1	0	0	0	1
Proceedings of 2015 IEEE International Conference on Research	0	1	0	0	0	0	0	1

in Computational Intelligence and Communication Network 2015								
Proceedings of the 2017 International Conference on Smart Technology for Smart Nation Smart 2017	0	0	0	1	0	0	0	1
Proceedings of the 2018 IEEE Symposium Series on Computational intelligence SSCI 2018	0	0	0	0	1	0	0	1
Proceedings of the ACM SIGKDD International Conference on Knowledge Recovery And Data Mining	1	0	0	0	0	0	0	1
Proceedings of the National Academy of Sciences of the United States of America	0	0	0	0	0	1	0	1
Recent Patents On Computer Science	0	0	0	0	1	0	0	1
Sustainable Computing Informatics And Systems	0	0	0	0	0	1	0	1

3.8 Statistics of Funding Sponsor

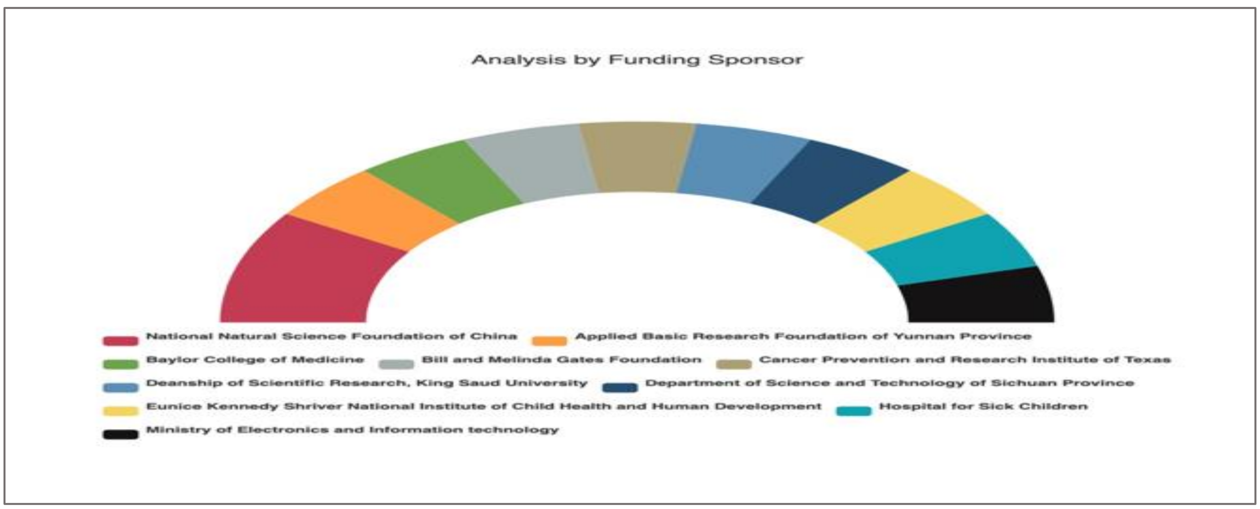

Fig. 32 Funding Sponsor

Source: Scopus_DB (accessed on 16th October 2020)

Figure 19 highlights the funding sponsors. The key sponsor for the publications is the "National Natural Science Foundation" of China.

4. Limitations of the present study

The bibliometric study presented in this research takes into account publications from the Scopus and Web of Science database only. There may be a possibility of a few other journals and publications present in databases like Google Scholar and PubMed, which have not been considered during the data analysis of this research. Thus, they have been excluded from the diagnosis completely. In addition to this, the calculation of citations has been drawn out from the Scopus database only. Different research databases show various statistics of citations. Along with this, the research is limited to the English language only.

5. Conclusion

The bibliometric examination enables researchers to gain a more in-depth insight into the topic's potential and identify the loops. This in a way helps to recognize the different variables that could be considered during research in feature Selection. The study is mainly based on the Scopus information base, which comprehended the imminent authors, publication, citation, and co-appearance among them and so on.

The keyword examination for the most part helps in choosing the further exploration. This paper primarily centers around executing feature selection using optimization techniques in healthcare so that readers will get idea about different aspects like influential source titles, authors, keywords, subject areas, publication types, their languages etc.

The study's findings reveal that research-oriented towards these subject areas could benefit and revolutionaries will make the health care field better. Optimized feature selection is the most essential and stimulate field of study and therefore it was necessary to tackle the detailed bibliometric study about this topic. That's why this paper is formulated.

References

1. Sowmiya, C., & Sumitra, P. (2020). A hybrid approach for mortality prediction for heart patients using ACO-HKNN. *JOURNAL OF AMBIENT INTELLIGENCE AND HUMANIZED COMPUTING*.
2. Manimaran, R., & Vanitha, M. (2018). Prediction of Diabetes Disease Using Classification Data Mining Techniques. *International Journal of Engineering and Technology (IJET), ISSN (Print)*.
3. Grewal, G. K., & Kaur, K. Heart Disease Prediction System using Binary Particle Swarm Optimization Algorithm.
4. Shaffiee Haghshenas, S., Pirouz, B., Shaffiee Haghshenas, S., Pirouz, B., Piro, P., Na, K. S., ... & Geem, Z. W. (2020). Prioritizing and Analyzing the Role of Climate and Urban Parameters in the Confirmed Cases of COVID-19 Based on Artificial Intelligence Applications. *International Journal of Environmental Research and Public Health*, *17*(10), 3730.
5. Sardinas, R. Q., Santana, M. R., & Brindis, E. A. (2006). Genetic algorithm-based multi-objective optimization of cutting parameters in turning processes. *Engineering Applications of Artificial Intelligence*, *19*(2), 127-133.
6. Izakian, Z., Mesgari, M. S., & Weibel, R. (2020). A feature extraction based trajectory segmentation approach based on multiple movement parameters. *Engineering Applications of Artificial Intelligence*, *88*, 103394.
7. Bu, F. (2018). An efficient fuzzy c-means approach based on canonical polyadic decomposition for clustering big data in IoT. *Future Generation Computer Systems*, *88*, 675-682.
8. Bharathi, R., Abirami, T., Dhanasekaran, S., Gupta, D., Khanna, A., Elhoseny, M., & Shankar, K. (2020). Energy efficient clustering with disease diagnosis model for IoT based sustainable healthcare systems. *Sustainable Computing: Informatics and Systems*, *28*, 100453.
9. Yeh, K. H. (2016). A secure IoT-based healthcare system with body sensor networks. *IEEE Access*, *4*, 10288-10299.
10. Gope, P., Gheraibia, Y., Kabir, S., & Sikdar, B. (2020). A secure IoT-based modern healthcare system with fault-tolerant decision making process. *IEEE Journal of Biomedical and Health Informatics*.

11. Surani, S., Sheta, A., Turabieh, H., Park, J., Mathur, S., & Katangur, A. (2019). DIAGNOSIS OF SLEEP APNEA USING ARTIFICIAL NEURAL NETWORK AND BINARY PARTICLE SWARM OPTIMIZATION FOR FEATURE SELECTION. *Chest*, *156*(4), A136.
12. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.
13. Gürüler, H. (2017). A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. *Neural Computing and Applications*, *28*(7), 1657-1666.
14. Majeed Alneamy, J. S., & Hameed Alnaish, R. A. (2014). Heart disease diagnosis utilizing hybrid fuzzy wavelet neural network and teaching learning based optimization algorithm. *Advances in Artificial Neural Systems*, *2014*.
15. Resende, P. A. A., & Drummond, A. C. (2018). Adaptive anomaly-based intrusion detection system using genetic algorithm and profiling. *Security and Privacy*, *1*(4), e36.
16. Pourpanah, F., Lim, C. P., Wang, X., Tan, C. J., Seera, M., & Shi, Y. (2019). A hybrid model of fuzzy min–max and brain storm optimization for feature selection and data classification. *Neurocomputing*, *333*, 440-451.
17. Li, H., Zhao, J., Zhang, X., & Ni, X. (2017). Fault Diagnosis for Machinery based on Feature Selection and Probabilistic Neural Network. *International Journal of Performability Engineering*, *13*(7).
18. Özşen, S. (2013). Classification of sleep stages using class-dependent sequential feature selection and artificial neural network. *Neural Computing and Applications*, *23*(5), 1239-1250.
19. Priscila, S. S., & Hemalatha, M. (2018). Diagnosis of heart disease with particle bee-neural network.
20. Hardie, R. C., Vaidyanathan, M., & McManamon, P. F. (1998). Spectral band selection and classifier design for a multispectral imaging laser radar. *Optical Engineering*, *37*(3), 752-762.
21. Tahir, M., Tubaishat, A., Al-Obeidat, F., Shah, B., Halim, Z., & Waqas, M. (2020). A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare. *Neural Computing and Applications*, 1-22.
22. Puyalnithi, T., & Vankadara, M. (2018). A Unified Feature Selection Model for High Dimensional Clinical Data Using Mutated Binary Particle Swarm Optimization and Genetic

Algorithm. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, *13*(4), 1-14.

23. Sayed, G. I., Hassanien, A. E., & Azar, A. T. (2019). Feature selection via a novel chaotic crow search algorithm. *Neural computing and applications*, *31*(1), 171-188.
24. Alirezaei, M., Niaki, S. T. A., & Niaki, S. A. A. (2019). A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. *Expert Systems with Applications*, *127*, 47-57.
25. Yahyaoui, A., & Yumuşak, N. (2018). Decision support system based on the support vector machines and the adaptive support vector machines algorithm for solving chest disease diagnosis problems.
26. Daliri, M. R. (2012). Feature selection using binary particle swarm optimization and support vector machines for medical diagnosis. *Biomedical Engineering/Biomedizinische Technik*, *57*(5), 395-402.
27. Khan, S., Khan, A., Maqsood, M., Aadil, F., & Ghazanfar, M. A. (2019). Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare. *Journal of Grid Computing*, *17*(2), 239-254.
28. Young-Min, K. (2019). Feature visualization in comic artist classification using deep neural networks. *Journal of Big Data*, *6*(1), 1-18.
29. Waseem, M. H., Nadeem, M. S. A., Abbas, A., Shaheen, A., Aziz, W., Anjum, A., ... & Shim, S. O. (2019). On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction. *IEEE Access*, *7*, 141072-141082.
30. Choi, H., Yeo, D., Kwon, S., & Kim, Y. (2011). Gene selection and prediction for cancer classification using support vector machines with a reject option. *Computational statistics & data analysis*, *55*(5), 1897-1908.
31. Selvakuberan, K., Indradevi, M., & Rajaram, R. (2008). Combined Feature Selection and classification–A novel approach for the categorization of web pages. *Journal of Information and Computing Science*, *3*(2), 083-089.
32. Ghamisi, P., & Benediktsson, J. A. (2014). Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *IEEE Geoscience and remote sensing letters*, *12*(2), 309-313.

33. Gupta, A., & Purohit, A. (2017). RGAP: A Rough Set Genetic Algorithm and Particle Swarm Optimization based Feature Selection Approach. *International Journal of Computer Applications*, *161*(6).
34. Jabor, A. H., & Ali, A. H. (2019). Dual heuristic feature selection based on genetic algorithm and binary particle swarm optimization. *Journal of University of Babylon for Pure and Applied Sciences*, *27*(1), 171-183.
35. Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, *6*, 29637-29647.
36. Ding, Y., Zhou, K., & Bi, W. (2020). Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer. *Soft Computing*, 1-10.
37. Nandhini, M., & Sivanandam, S. N. (2015). An improved predictive association rule based classifier using gain ratio and T-test for health care data diagnosis. *Sadhana*, *40*(6), 1683-1699.
38. Baranidharan, T., Sumathi, T., & Chandra Shekar, V. (2016). Weight optimized neural network using metaheuristics for the classification of large cell carcinoma and adenocarcinoma from lung imaging. *Current Signal Transduction Therapy*, *11*(2), 91-97.
39. Hu, Z., Bao, Y., Xiong, T., & Chiong, R. (2015). Hybrid filter–wrapper feature selection for short-term load forecasting. *Engineering Applications of Artificial Intelligence*, *40*, 17-27.
40. Higuchi, K., Sato, K., Makuuchi, H., Furuse, A., Takamoto, S., & Takeda, H. (2006). Automated diagnosis of heart disease in patients with heart murmurs: application of a neural network technique. *Journal of medical engineering & technology*, *30*(2), 61-68.
41. Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. *IEEE access*, *4*, 9145-9154.
42. Baranidharan, T., Sumathi, T., & Chandra Shekar, V. (2016). Weight optimized neural network using metaheuristics for the classification of large cell carcinoma and adenocarcinoma from lung imaging. *Current Signal Transduction Therapy*, *11*(2), 91-97.