Diabetes Prediction using Machine learning : A Bibliometric Analysis

Sina Patel[1], Vijayshri Nitin Khedkar[2]
[1]sinapatel05@gmail.com, [2]vijayshrik17@gmail.com
[1,2] Symbiosis Institute of Technology, Symbiosis International [Deemed University], Mulshi, Maharashtra - 412115

ABSTRACT

Diabetes is a chronic disease which can be deadly if undetected. Artificial intelligence is helping in healthcare industry to a great extent. Diabetes, if predicted at an early stage can help many people to save lives. Decision-making, diagnosing and predicting diabetes have become an increasing trend in recent years. There are numerous publications in diabetes prediction and yet it's an ongoing research topic with availability of new data and methods. This study aims to provide a global picture on current status of diabetes prediction research field by analyzing trends in publications, authors, countries and institutions from Scopus and Web of Science database. Further, to enhance the analysis, network inspection in terms of co-authorship, collaborative countries, citation analysis and keyword co-occurrences were explored.

1. INTRODUCTION

Diabetes is deadly disease if left undiagnosed. It is characterized as a chronic disease which affects the level of glucose in blood. There are about 422 million people living with diabetes majority falling under low and middle income groups. Also, there are more than a million deaths directly related to diabetes every year (1). Diabetes causes serious damage to eyes, heart, kidneys

and nerves which affect human functioning. There are mainly two types of diabetes Type 1 and Type 2 diabetes. Type 1 occurs when there is lack of insulin in body which permanently damages cells that produce pancreas. The second type of diabetes is when the body cannot effectively use the insulin produced by it. The exact cause is not identified; however genetics and lifestyle may contribute as risk factors. There is another type which occurs during pregnancy called gestational diabetes. While diabetes in general is not curable, but with proper medications it can be controlled.

Diabetes Mellitus is diagnosed through a professional doctor or an automated device. However, often symptoms of diabetes at initial phase are so minute which often a specialist might not recognize it (2). As a result, machine learning and artificial intelligence is proving to be efficient way to diagnose or predict diabetes as early as possible.

Advantages of predicting early diabetes are decreased risk to human health and workload to medical specialists. Many researchers have used data mining and machine learning algorithms to predict diabetes. Factors like blood sugar, BMI, skin thickness, age affects the prediction of diabetes(3).

The bibliometric study helps to know current progress of in research work done in any particular field by analyzing the literature in the form different articles, conference papers, etc which are available across various database. Using different statistical tools, different trends and developments can be tracked. Tools like Tableau and VOSviewer were used for analysis.

Aim of this bibliometric study:

a. To identify amount of research work done in the field of diabetes prediction for the year 2009-2020.

b. To perform citation analysis and identify contributing authors.

c. To identify which country is productive in this field.

d. To identify the directions of publications based on affiliated institutions and sponsors.

Section 2 presents the research methodology: how data was collected and analyzed. Section 3 has statistical information followed by section 4 which shows network analysis and citation analysis. In Section 5, discussions from analysis and limitation of the study are shown and last section 6 gives conclusion.

2. RESEARCH METHODOLOGY

Researchers should have a good knowledge about on-going research and the information about the authors, institutions, countries who are contributing to the research. Webometric, scientometric, h-index and bibliometric are few such methods are used for analyzing the trends(4). In this study, bibliometric analysis is done which includes publication types, geographical information, keywords and authors for analyzing the researchers. There are several databses available which contains huge amount of data related to published papers such as Web Of Science, IEEE, Google Scholar, Scopus and many more. In this paper, we have primarily used Scopus and Web of science (WoS) which contains massive amount of data.

2.1 **Significant Keywords**

The keyword strategy for both database: Scopus and Web Of Science(WoS) is classified into primary keywords and secondary keywords which is shown in table for the year 200-2020.

Primary Keywords	AND	“Diabetes”
		“prediction”
Secondary Keywords	OR	“Machine learning”
		“Data mining”
		“Neural networks”
		“Artificial Intelligence”

Table 1 Keywords Used

2.2 **Initial Search Results**

No.	Language	No. of Documents
1	English	2190
2	Chinese	14
3	Persian	11
4	Russian	3
5	Spanish	2
6	Arabic	1
7	German	1
8	Portuguese	1
9	Turkish	1
Total		2224

Table 2 No. of Documents by language (www.scopus.com, accessed on 04-11-2020)

No.	Language	No. of Documents
1	English	670
2	Korean	4
3	German	1
4	Spanish	1
5	Chinese	1
Total		677

Table 3 No. of Documents By language (WoS)

Initial search through planned keywords generated 2224 publications and 677 publications for Scopus and WoS database respectively. Then it is restricted to

documents in English only, which is 2190 and 677 documents (table). All kinds of papers: published or unpublished are considered for statistical analysis.

3. STATISTICAL INFORMATION

3.1 Publication trends

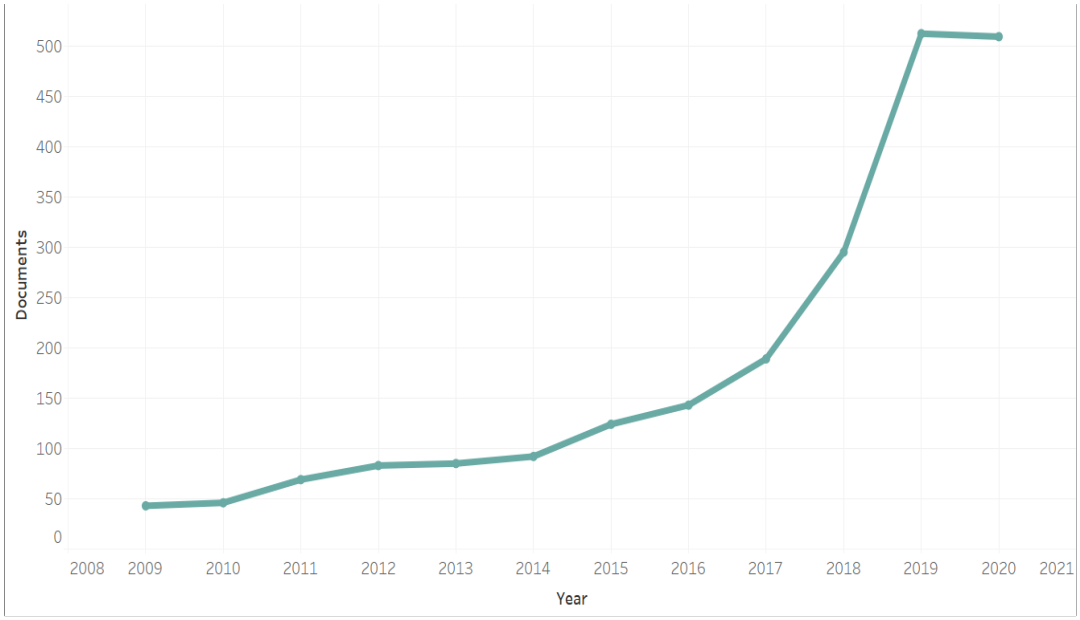

Figure 1 Publications by Year (www.scopus.com, accessed on 04-11-2020)

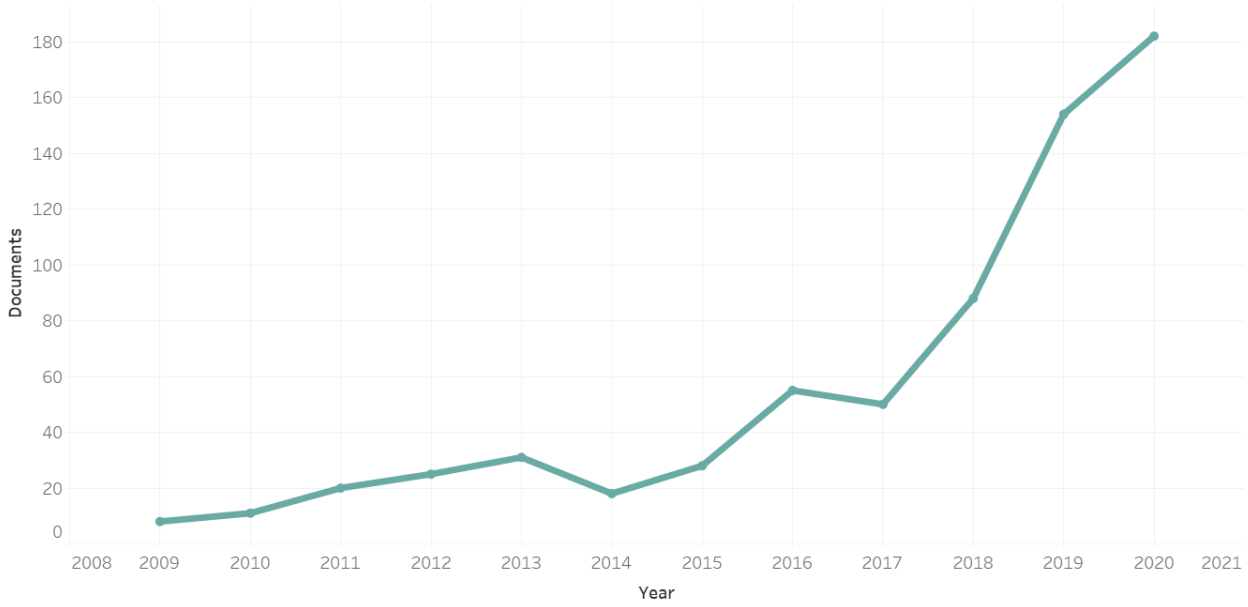

Figure 2 Publication by Year (WoS)

Figure 1 and Figure 2 shows yearly publication of diabetes prediction from the year 2009 to 2020 for Scopus and WoS database respectively. It can be observed that there is a gradual increase in number of publications. Moreover, it can be seen that between the years 2017-2019 the research interest in the field has increased rapidly. The consistent growth in both databases shows research opportunities and advancements in predicting diabetes.

3.2 Type of document

It can be derived that nearly 60% of documents are articles from both database as seen in figure 4 and figure 4. In Scopus, conference paper have nearly 30% of total publications as shown in figure 3 and about 20% of papers were kept under category of others in WoS (figure 4). Whereas, 4% paper were review type in Scopus as well as WoS.

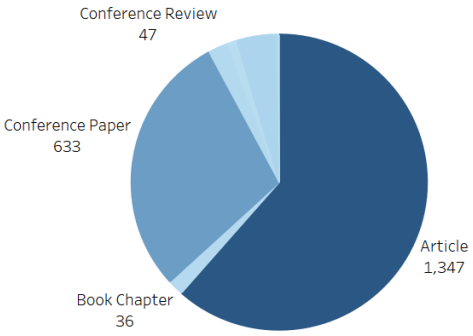

Figure 3 Documents by type (www.scopus.com, accessed on 04-11-2020)

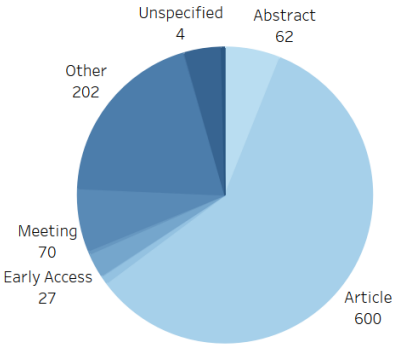

Figure 4 Documents by type (WoS)

3.3 Source statistics

Figure 5 shows source statistics for diabetes prediction publications from Scopus database. Statistics shows that maximum numbers of publications are from Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, Plos One, Advances in Intelligent Systems and Computing, etc

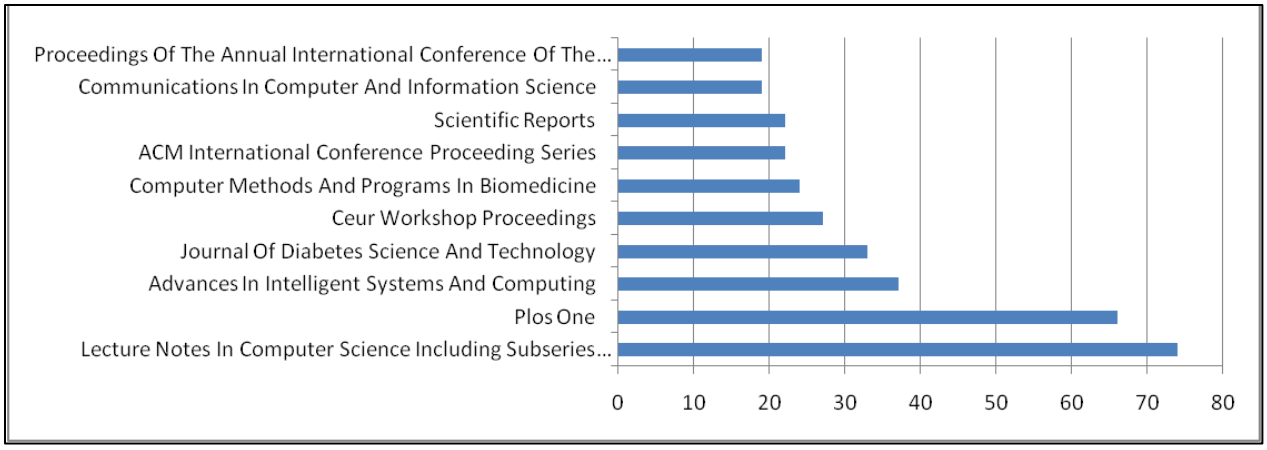

Figure 5 Top 10 sources (www.scopus.com, accessed on 04-11-2020)

Figure 6 shows the number of documents published by top 10 sources in WoS. It can be clearly seen PLOS ONE published highest number of documents followed by diabetes. Also, PLOS One and scientific reports were both common in top 10 sources of both databases.

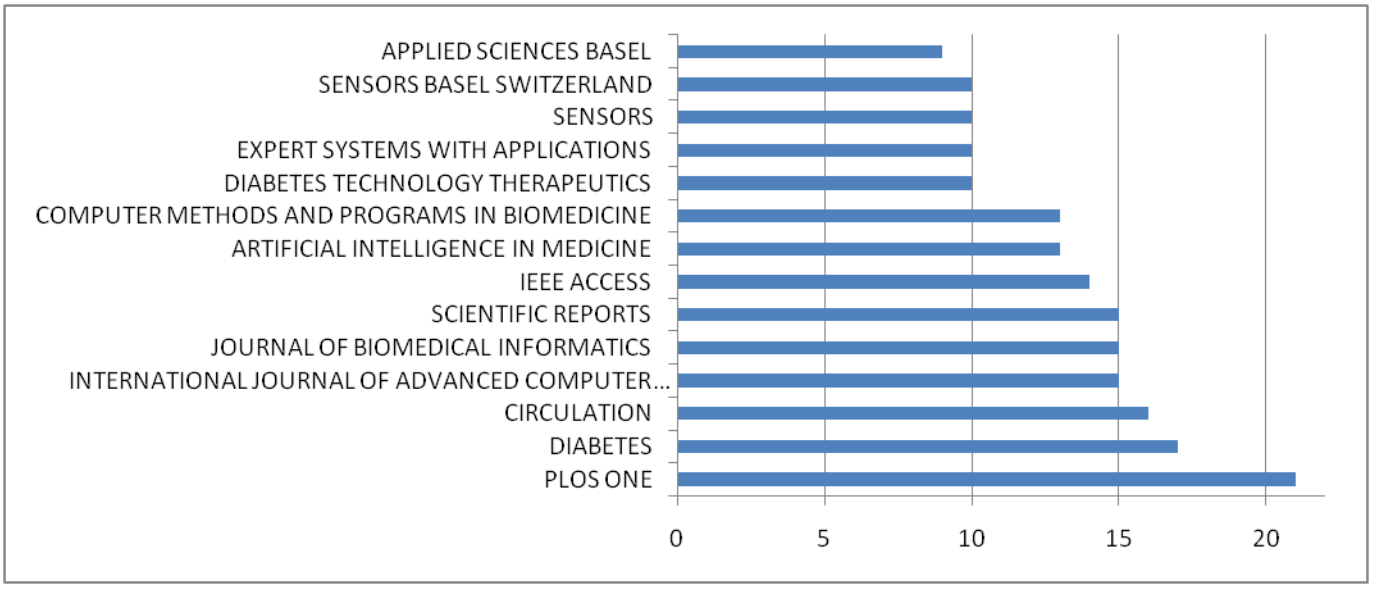

Figure 6 Top 10 sources (WoS)

3.4 Prominent Subject Areas

Figure shows specific information on subject areas for diabetes prediction publications extracted from Scopus and WoS database. The maximum number of publications are concentrated in Computer Science(27%), followed by Medicine (22%) and then engineering at 18% for Scopus. Whereas, WoS has 16% of documents in mathematical computational biology and computer each, followed by 14% mathematics papers.

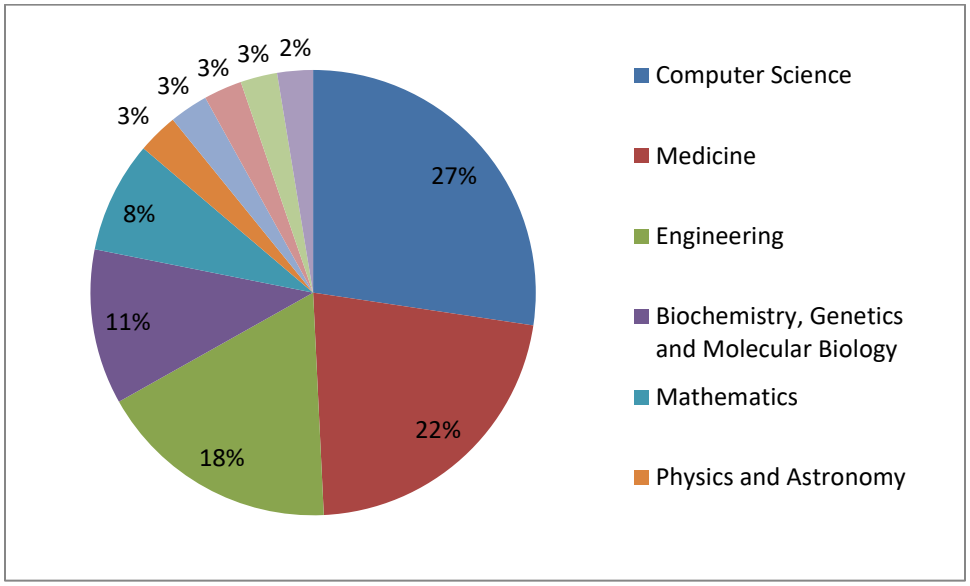

Figure 7 Documents by subject area (www.scopus.com, accessed on 04-11-2020)

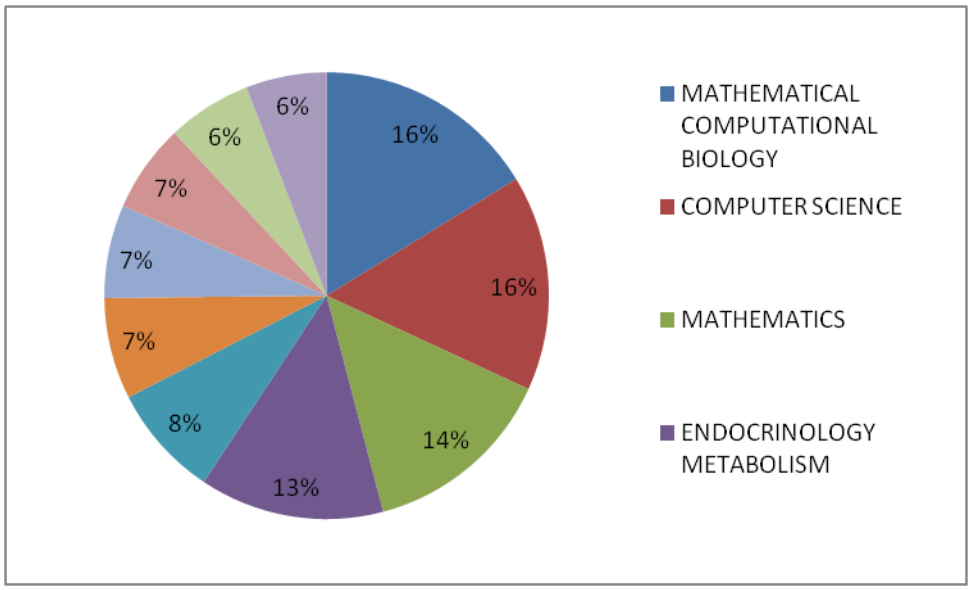

Figure 8 Documents by subject area (WoS)

3.5 Funding Sponsors

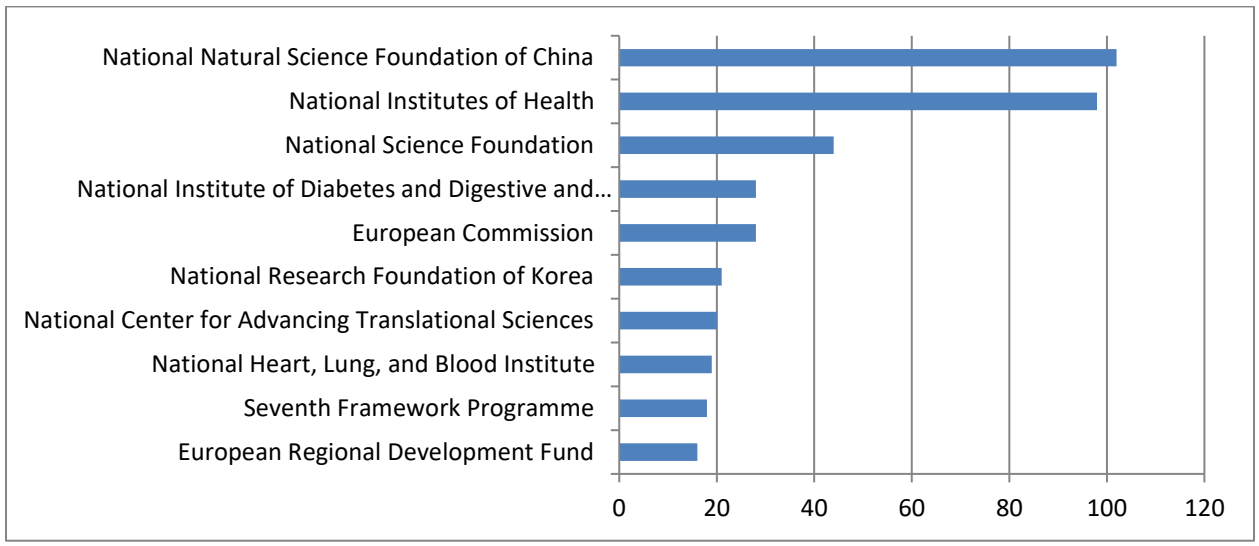

Figure 9 Top Funding Sponsors (www.scopus.com, accessed on 04-11-2020)

Figure 9 shows top ten agencies funding diabetes prediction. National Natural Science Foundation of China is important fund provider followed by National institute of Health in the

field of diabetes prediction. It is worth noting that researchers are taking lead for development of this research area by availing funding through funding agencies.

3.6 Countries

Table 4 shows top 10 contributing countries in diabetes prediction research in Scopus Database. It is clearly inferred that Unites states of America is leading with 28% of publications followed by Indians with 24% and Chinese with 16%.Table 5 shows top 10 countries having publications in the area of diabetes prediction from WoS database. Undoubtedly, USA has maximum publications with 28% of total publications followed by People's Republic of China with 15% and India having 10% of publications. However, the order of top publishing countries are different from in both table 4 and table 5, the prominent countries remain same.

Country	Percentage
United States	28%
India	24%
China	16%
United Kingdom	7%
Germany	5%
Spain	5%
Italy	5%
South Korea	4%
Australia	4%
Canada	3%

Table 4 Top 10 Countries (www.scopus.com, accessed on 04-11-2020)

Country	Percentage
USA	28%
People's R China	15%
India	10%
England	6%
South Korea	6%
Italy	5%
Iran	5%
Canada	4%
Spain	4%
China	4%

Table 5 Top 10 countries (WoS)

Figure 10 and figure 11 shows map visualization of countries contributing from Scopus and WoS respectively. The darker the shade of country, more the number of publications.

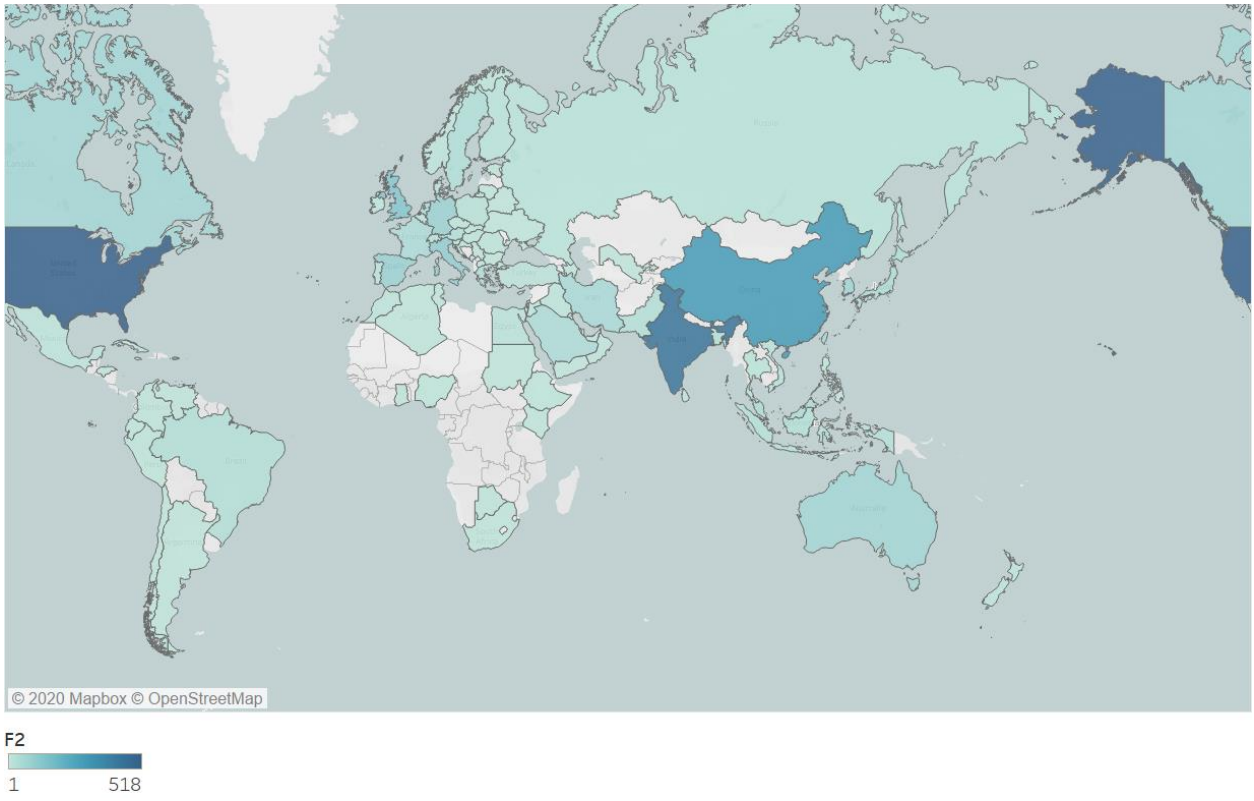

Figure 10 Map Visualization of contributing countries (www.scopus.com, accessed on 04-11-2020)

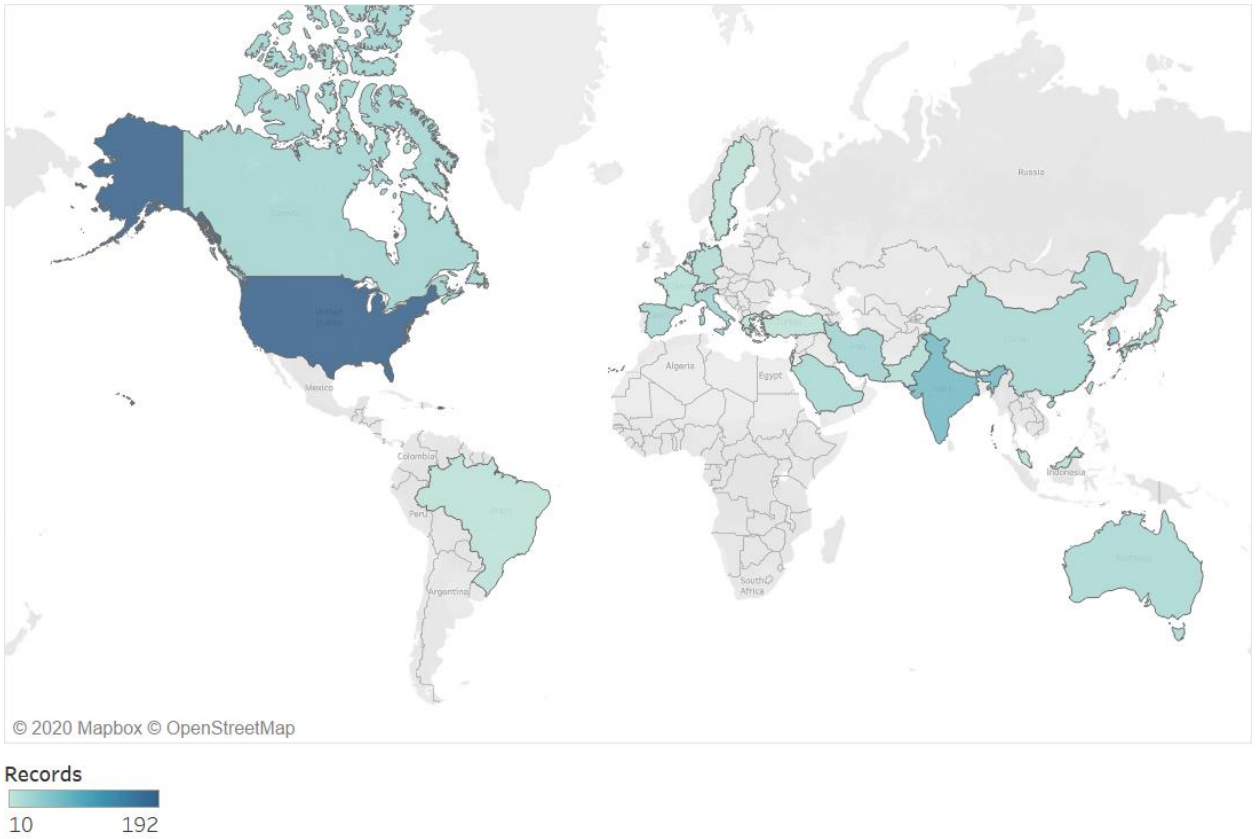

Figure 11 Map Visualization of Contributing countries (WoS)

3.7 Institutions

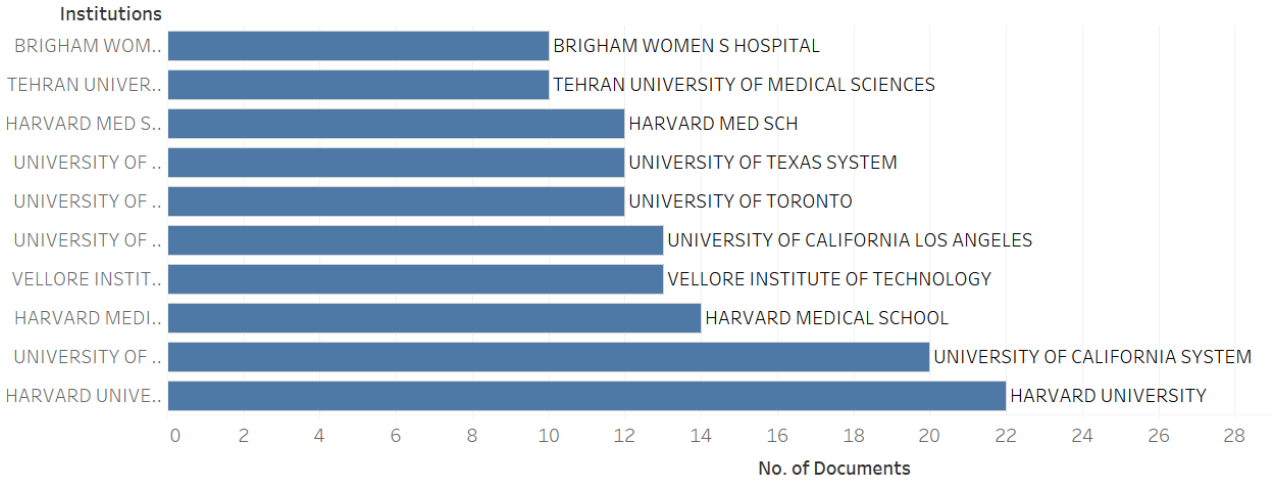

Figure 12 Top affiliated institutions(WoS)

Figure 12 shows top 10 affiliated organizations in the field of diabetes prediction. Harvard University sponsored highest amount of studies as per Web of Science Database. Also, institutions like University of California, Vellore institute of technology are also contributing in this field.

4. NETWORK ANALYSIS

Network analysis helps researchers to identify various members of network and relationship between them and the most important member of the network. Here, analysis of collaborative co-authors, collaborative countries, top cited documents and keyword co-occurrence analysis of articles which are at final published stage in Scopus database is presented as per data availability. This will give insight about which countries and authors are contributing in diabetes prediction filed and collaborating,

4.1 Collaborative Co-Authors Network Analysis

Co-authors network analysis helps researchers to understand insights of important authors of a particular field. Figure 13 shows co-authorship collaborative network of diabetes prediction using VOSviewer tool. Different authors are classified by different clusters presented in different colors in the visualization. The size of the circle shows the number of documents. Larger the size, more the number of documents. The minimum number of documents of author to be included in visualization is set to 5. As shown in figure, there are total 8 clusters and 210 links. In the field of diabetes prediction, the largest contributor of co-authors is Wang y. as shown in figure. The collaborative relationship between two authors is shown by connecting lines. Clusters having same color show strong cooperative relationship between authors. For example, main core author of red cluster is Li has stronger cooperative relationship than Liu in sky blue cluster.

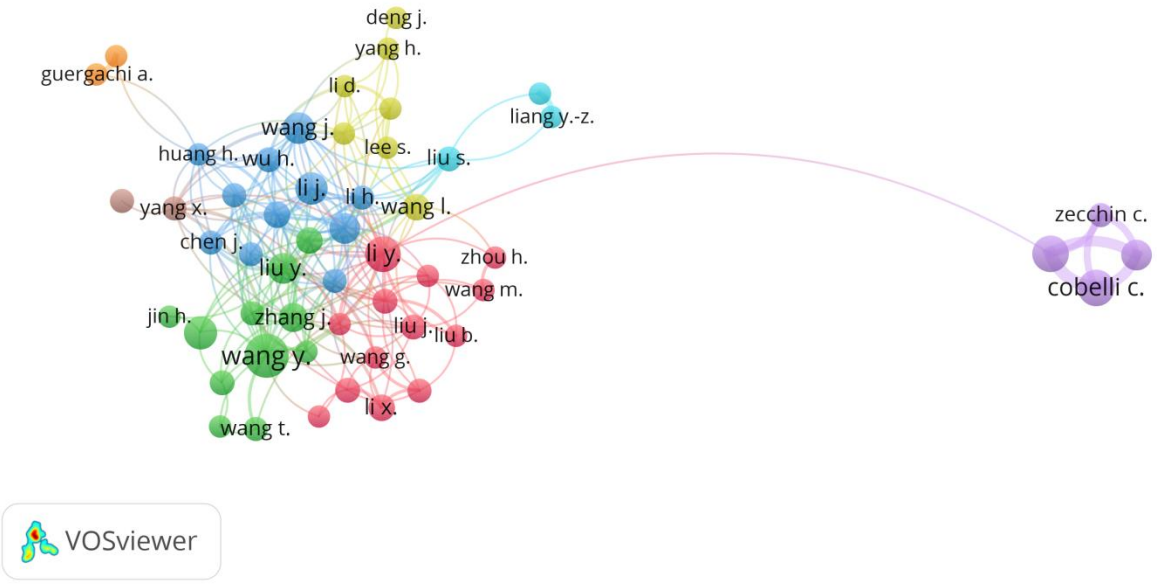

Figure 13 Collaborative author network analysis

4.2 Collaborative Countries Network Analysis

According to Scopus downloaded data, documents were derived from 96 countries. Out of which 46 countries met our threshold 5 documents per country, which were then considered for visualization in VOSviewer tool. Larger the size of cluster more the number of publication similar to figure 13. 6 clusters were formed as shown in figure 14. Thicker the line between countries, stronger the cooperative relation between them. For example, USA has stronger collaborative relationship with Canada than Iran since USA and Canada has thicker line than USA and Iran as seen in figure.

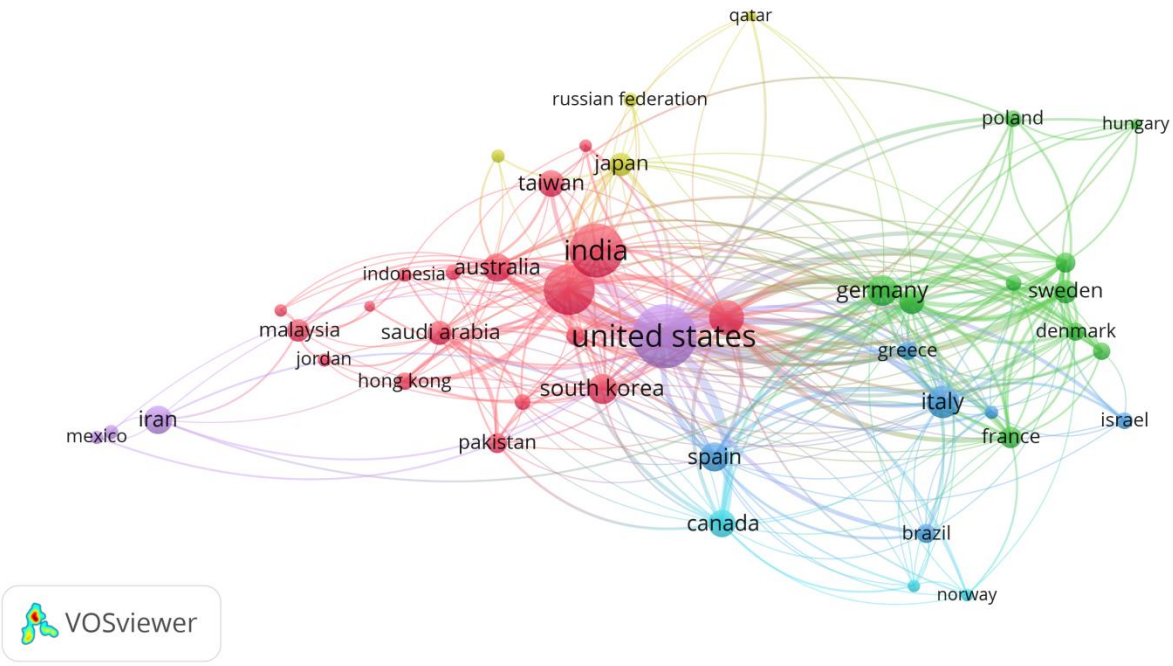

Figure 14 Collaborative country Network analysis

Country	Documents	Citations	Total link strength
United states	367	6053	276
India	241	1186	37
China	206	2500	128
United kingdom	82	1266	146
Italy	68	1229	66
Germany	58	1312	104
South korea	56	625	40
Spain	51	557	49
Iran	49	556	12
Australia	48	1200	51

Table 6 Top 10 countries collaborating

It is notably seen that USA has highest number of citations. Also, its link strength is highest among all countries which shows that it is most productive country in diabetes prediction. Moreover, due to its reputation research experts are interested in cooperating with the USA. On the one hand, it is probable that the USA has always been a global scientific leader because of

the scale of its economy and the level of its research effort(5). It will be worth noting that though India stands second in number of publication its link strength and citation count are comparably low than other countries.

4.3 Keyword Co-occurrence Analysis

The basic information like methods, goals about any article can quickly be obtained from keywords. Keyword co-occurrence is defined as when two or more keywords appear in the same article at the same time. Out of 1308 documents, 130 keywords were extracted using threshold of minimum 5 times occurrence of VOSviewer and the visualization can be seen in figure. The size of circle is directly proportional to the number of occurrence of keyword. The connecting lines indicate the strength between the keywords. There are total 9 clusters with 9 different colors which denote different categories as seen in figure 15. For example, green group mainly presents the machine learning methods used for predicting diabetes, red group indicates mainly medical terms related to risk factors. The insights of these keywords can help beginners to search related papers for researching.

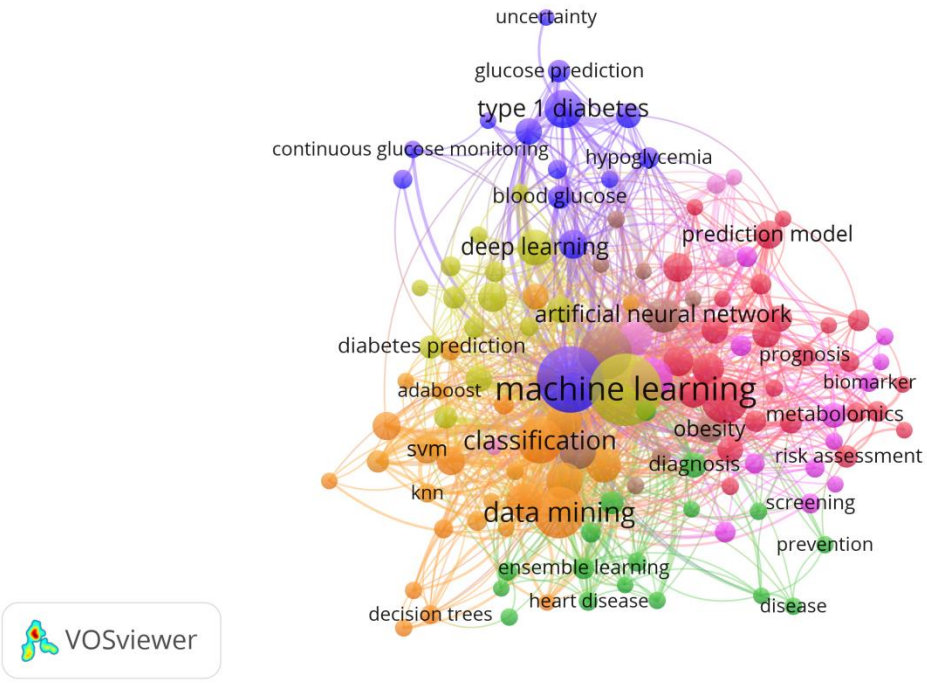

Figure 15 Keyword Co-occurrence

Alternatively, we have also listed top 20 keywords in table 7. It can be seen that words like support vector machine, random forest are also mentioned in top 20 lists. Keywords have interesting information, for example machine learning, neural networks and data mining. In the research of disease prediction the main objective is to improve the accuracy considering varied risk factors. It can be derived as these methods are widely used for prediction purpose of diabetes. Also, electronic health records can be considered based on real life situation to predict diabetes in improvised manner to enhance the accuracy of model. Thus, it is important to build effective predictive model which predicts diabetes at early stage without advanced lab tests.

Rank	Keyword	Occurrences	Total link strength	Rank	Keyword	Occurrences	Total link strength
1	machine learning	196	376	11	deep learning	34	64
2	Diabetes	160	330	12	feature selection	34	69
3	Prediction	92	208	13	logistic regression	33	87
4	data mining	85	169	14	artificial neural network	32	61
5	type 2 diabetes	65	107	15	random forest	32	74
6	Classification	64	142	16	artificial intelligence	29	75
7	diabetes mellitus	50	83	17	disease prediction	26	54
8	type 1 diabetes	41	62	18	electronic health records	22	34
9	decision tree	36	74	19	risk prediction	22	36
10	support vector machine	36	83	20	neural network	21	38

Table 7 Top 20 keywords

4.4 Top Cited Articles

To measure the impact of any paper, citation analysis can be used which indicates the value of an article. Table 8 shows top 5 cited articles retrieved from Scopus database. This gives insight that "Personalized Nutrition by Prediction of Glycemic Responses"(6) was the most cited article and of the 60 documents considered for the h-index, the h-index is 58.

		<2016	2016	2017	2018	2019	2020	subtotal	>2020	total
Publication Year	Document Title	356	219	329	402	512	390	1852	4	2212
2016	Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records	0	4	68	124	196	108	500	1	501
2015	Personalized Nutrition by Prediction of Glycemic Responses	4	91	140	161	173	188	753	1	758
2014	A hybrid intelligent system for medical data classification	30	20	31	26	37	21	135	0	165
2012	Novel biomarkers for pre-diabetes identified by metabolomics	127	64	50	54	60	41	269	1	397
2010	Chemometrics in metabolomics-A review in human disease diagnosis	195	40	40	37	46	32	195	1	391

Table 8 Top 10 cited documents

CONCLUSION

Through thorough bibliometric analysis using available tools that many authors have already worked in the domain of diabetes prediction. Many institutions and sponsors are affiliating works related to this research field. With new techonology,methods and data collected from various sources such as electronic health records, survey data, smart wearable and many more can be utilized for predicting diabetes at early stage with easy to measure features such as BMI and lifestyle choices. Based on downloaded data from Scopus and Web of science for diabetes prediction domain from the year 2009-2020 the statistical analysis was performed. Futhermore, collaborative network analysis, co-authorship analysis and keyword analysis was performed.

According to above bibliometric analysis, following conclusions can be made as follows : (i) number of publications have gradually increased anually for the year 2009-2020;(ii) there are various different types of publications, but majority type of publications are articles; (iii) PLOS One and scientific reports are two common sources from top 10 sources of both Scopus and WoS; (iv) the most prominent subject area in this field is Computer science; (v) the top funding organizations is National natural science foundation of china (vi) the United states of America is most productive country in terms of publications, citations and co-authorship; (vii) though India has more publications but its collaborative scores less when compared to other countries in the list of top 10 countries; (viii) the most frequently cited paper is "Personalized Nutrition by Prediction of Glycemic Responses"(6) and these most frequently cited papers have high impact. These insights can be beneficial for some researchers who are interested in researching in the field og diabetes prediction to have basic knowledge and research characteristics.

There are some limitations of this study. We only utilized Scopus and Web of science for statistical analysis for the year 2009-2020 which can further be enhanced.

© 2021. This work is published under https://creativecommons.org/licenses/by-nc/(the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.