Spring 2-25-2021

Bibliometric Survey on Predictive Analysis using Multiple Regression for Invoice Generation System

Chetana Nadekar
Pimpri Chinchwad College of Engineering, Pune, India, nadekarchetana7@gmail.com

Subodh Awate
Pimpri Chinchwad College of Engineering, Pune, India, subodh.awate@gmail.com

Dattatray Damse
Pimpri Chinchwad College of Engineering, Pune, India, damsednd1111@gmail.com

Smita Khairnar
Pimpri Chinchwad College of Engineering, Pune, India, chavansmita31@gmail.com

Bibliometric Survey on Predictive Analysis using Multiple Regression for Invoice Generation System

Chetana Nadekar[1], Subodh Awate[2], Dattatray Damse[3], Smita Khairnar[4]

[1,2,3]Dept. of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India.

Email:
[1]nadekarchetana7@gmail.com, [2] subodh.awate@gmail.com, [3]damsednd1111@gmail.com

[4]Assistant Professor, Dept. of Computer Engineering ,Pimpri Chinchwad College of Engineering, SPPU,Pune, India
Email: [4]chavansmita31@gmail.com

ABSTRACT

The research is exploring the opportunities available for the application of Predictive analysis using multi regression for invoice generation systems. We work by analyzing the data generated and saved in the database of the clinic. This database contains the financial operations along and also contains other useful information that is being gathered about the patients.
We are proposing a different approach for a billing system to make the task easier for the generation of invoices for the clinic. The mapping of the relational data and using a predictive model that is based on a regression analysis that will predict the future values involved in the business. We found that predicting the future scope of the treatments can be done using a multi-regression model.

Keywords: Predictive Analysis, Invoice Generation, Multi Linear Regression

1. INTRODUCTION

In today's world, people are inclined more towards technology for the little things in life. The service and applications that are being offered today are increasingly interconnected. Various things are being developed using different available technology. Many things have become easier to do by using technology. Tasks that Humans find difficult are taken over by the technology for performing. And the precision has increased tenfold than Humans.

The Invoice Generation System is such a kind of software that will replace the tedious task of generating Invoices and doing billing using the traditional pen and paper approach. The purpose of this system is to replace the traditional way of calculating and generating invoices with a computer-based system. This approach was moderate, error-prone, and repetitive. It also takes time to analyze the generated invoices and conclude if the business is in profit or not. Various things can be done after analyzing the data generated through Invoices. These things could be helpful and profitable for the business of the clinic.

The main function of the system is to generate the Invoices for the clinics, collect and store the data in a single place, perform data analysis on these generated invoices, and predict the business needs accordingly. The Invoices will be generated at different locations and the data will be collected over the cloud in a single space. The storage for the data is Amazon Cloud S3 storage. This will give easy access to the User working from different locations. The data can be easily accessed and retrieved for processing as per the requirements. The storage of data over the cloud gives security for the system as well. The system will become fault-tolerant and there will be no downtime at the locations where the system will be used. To achieve predictive goals in the multi regression model, we found it useful in this system.

2. PRELIMINARY DATA COLLECTION

A bibliometric survey gives more insight into a particular topic. To meet this goal, Scopus Database is accessed by the library portal or one can access it by using an individual's login credentials. There are different databases. These databases are categorized into two parts: Open Access Database and Paid Database. Scopus database accessed in January 2021 has been considered in this paper as it is the largest among the peer-reviewed databases. A list of the keywords used is given in the following section.

2.1 Momentous Keywords

The momentous keywords required to search were "Predictive Analysis" and "Multi Regression". Secondary keywords related are "Invoice Generation" or "Billing System".

Search Query is formulated like:

Predictive Analysis” and “Multi Regression" or “Invoice Generation” or “Billing System” and PUBYEAR > 2009 and PUBYEAR < 2022 and (LIMIT-TO (EXACT KEYWORD, "Regression Analysis"))

Table 1: Planned search tactic for Keywords

Primary Keywords (AND)	**“Predictive Analysis” and “Multi regression”**
Secondary Keyword (OR)	**“Invoice Generation” or “Billing System”**

2.2 Initial Search Results

The number of publications generated for the Initial Search Query is 829. These publications are in different languages like English, Chinese, and German. Table 2 shows the trends in publishing languages. The majority number of publications in the English Language only.

Table 2: Trends in publishing language

Publication Type	**Publication Count**
English	796
Chinese	33
German	1
Total	829

Source: http://www.scopus.com (January 2021)

Research has been published in journal papers, reviews, articles, book chapters, conference proceedings, etc. for this survey. The researchers have publicized recent papers in Journal articles. 76.1% of journal articles and conference papers of 22.6 % were there. (Figure 1).

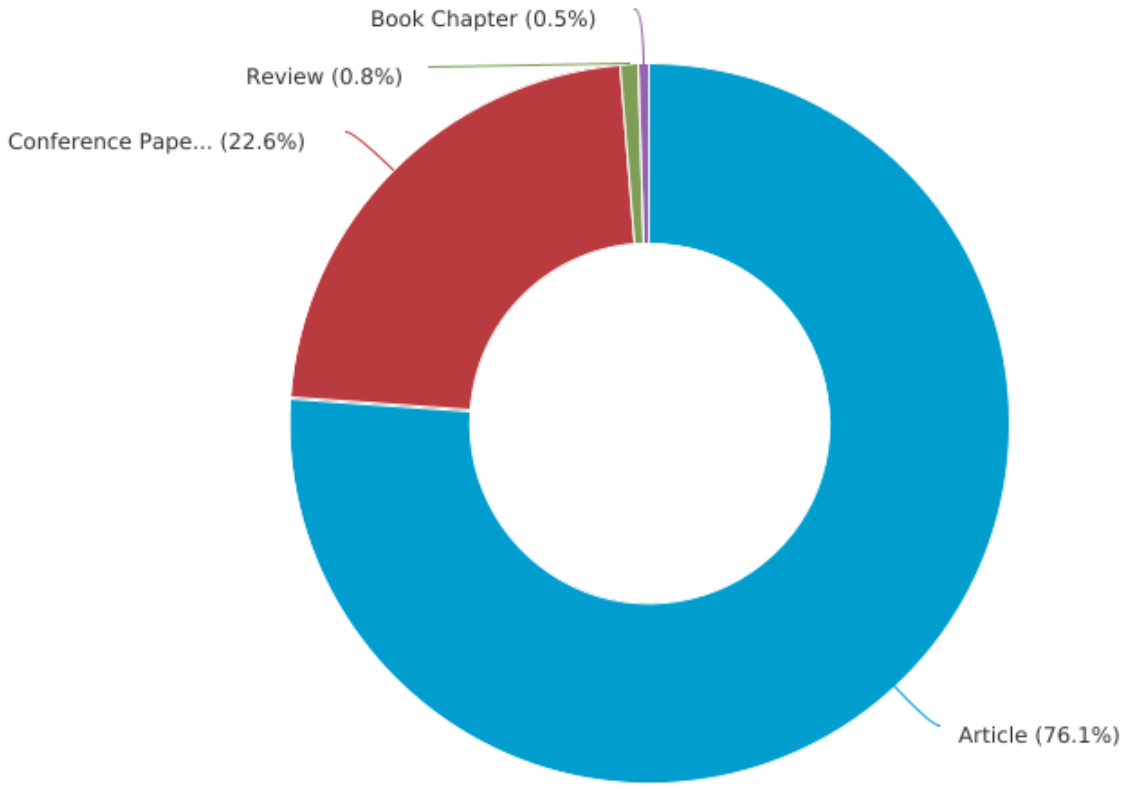

Figure 1: Analysis by Publication Type

Source: http://www.scopus.com (January 2021)

2.3 Highlights of Elementary data

The initial investigation is done based on the keywords that extracted 829 different types of publications from 2010 to 2020 in the area of predictive analysis using multi regression. The number of paper publications done per year is shown in table 3. The analysis is based on the number of publications per year as shown in figure 2. In 2019, the majority of researchers published their work.

Table 3: Publishing count per year

Year	Publication Count	Year	Publication Count
2021	19	2015	56
2020	98	2014	52
2019	118	2013	57
2018	80	2012	62
2017	86	2011	72
2016	74	2010	55

Source: http://www.scopus.com (January 2021)

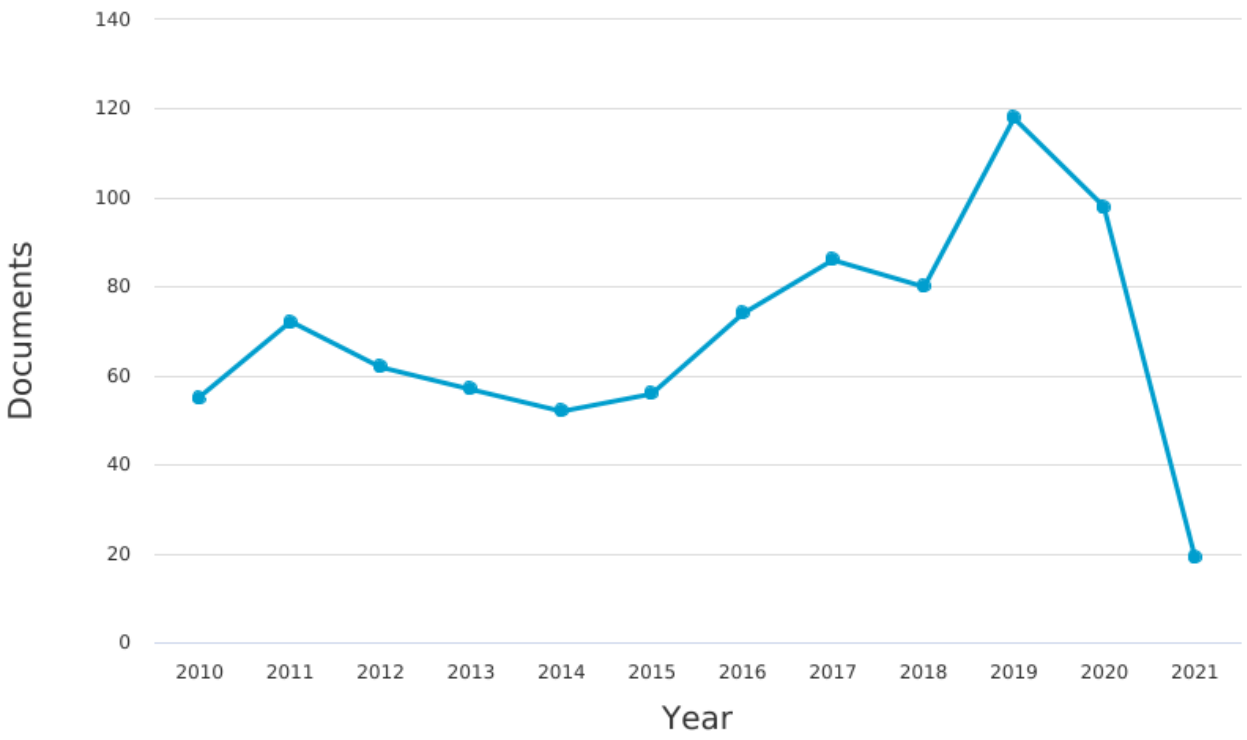

Figure 2: Analysis of Publication by Year

2.4 Geographical Region Analysis

The geographical region analysis of the published papers is shown in figure 3. The United States and Chinese publications are of the maximum number. Figures 3 and 4 show the contribution in publications by different countries. The United States has contributed up to 245 documents to this topic followed by the Chinese who contributed 170 documents.

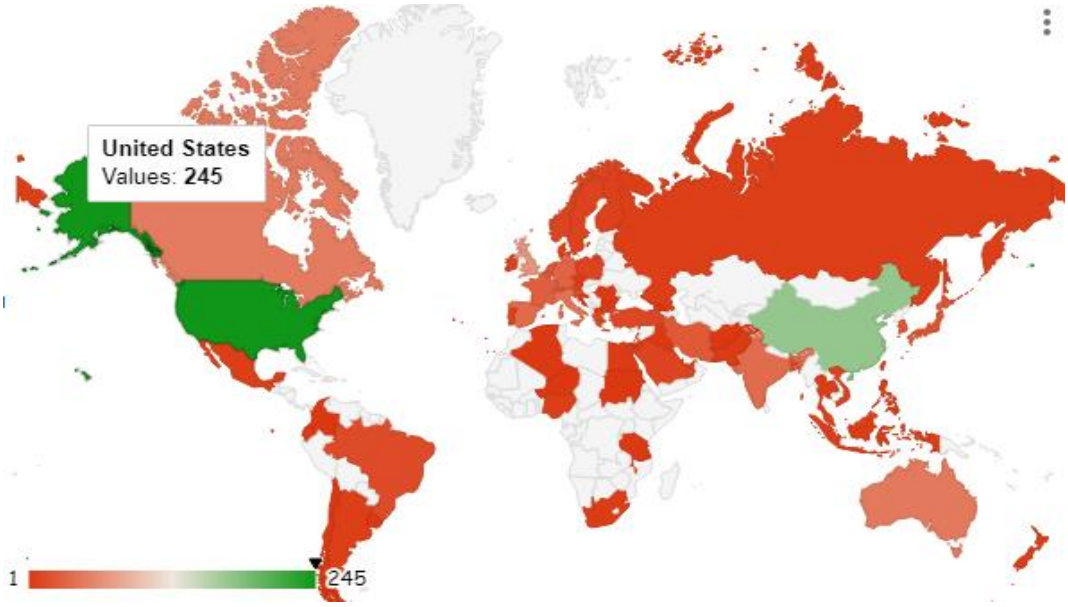

Figure 3: Research on Predictive Analysis using Multi Regression System at different geographic locations

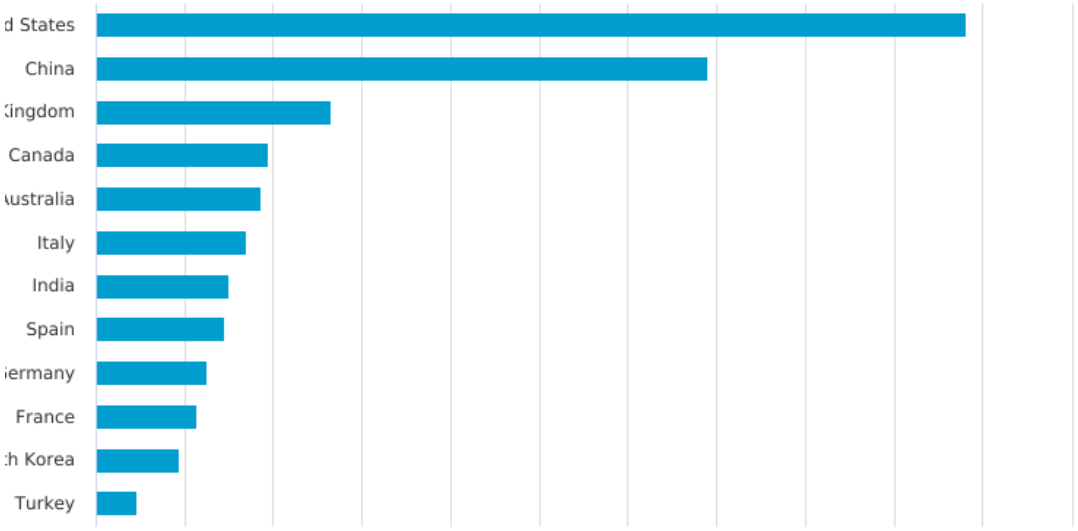

Figure 4: Analysis by country /Region

2.5 Analysis based on the subject area

This analysis discloses that a maximum number of research papers are published from the area of Computer Science, Engineering followed by Mathematics. The amount of research that is being carried out in different areas like Computer Science, Medicine, Engineering, and other fields is shown. Figure 5 shows the analysis of all fields that research about this topic and publish papers about it.

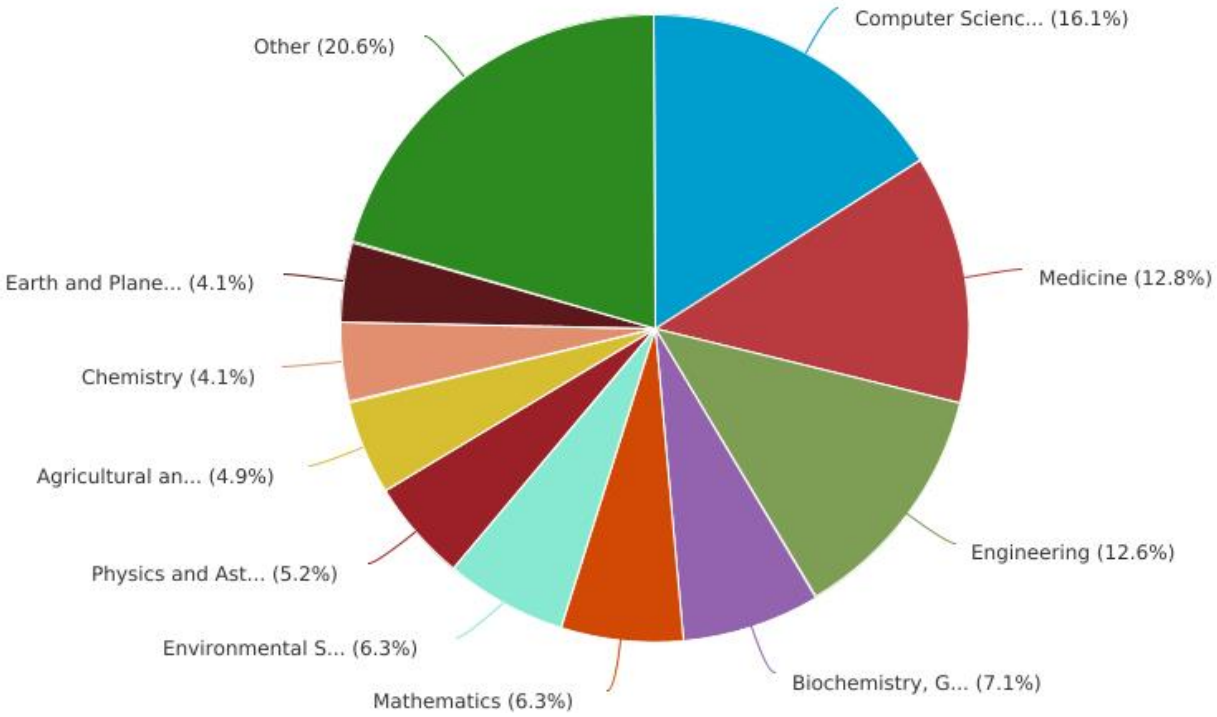

Figure 5: Analysis of Publications by Subject Area
Source: http://www.scopus.com (January 2021)

2.5 Analysis based on affiliation

Figure 6 shows an analysis of different Universities worldwide that contributed to publishing research work in this field. Jozef Stefan Institute has given a major contribution as shown in the figure. The top eleven universities publishing in this field have been shown.

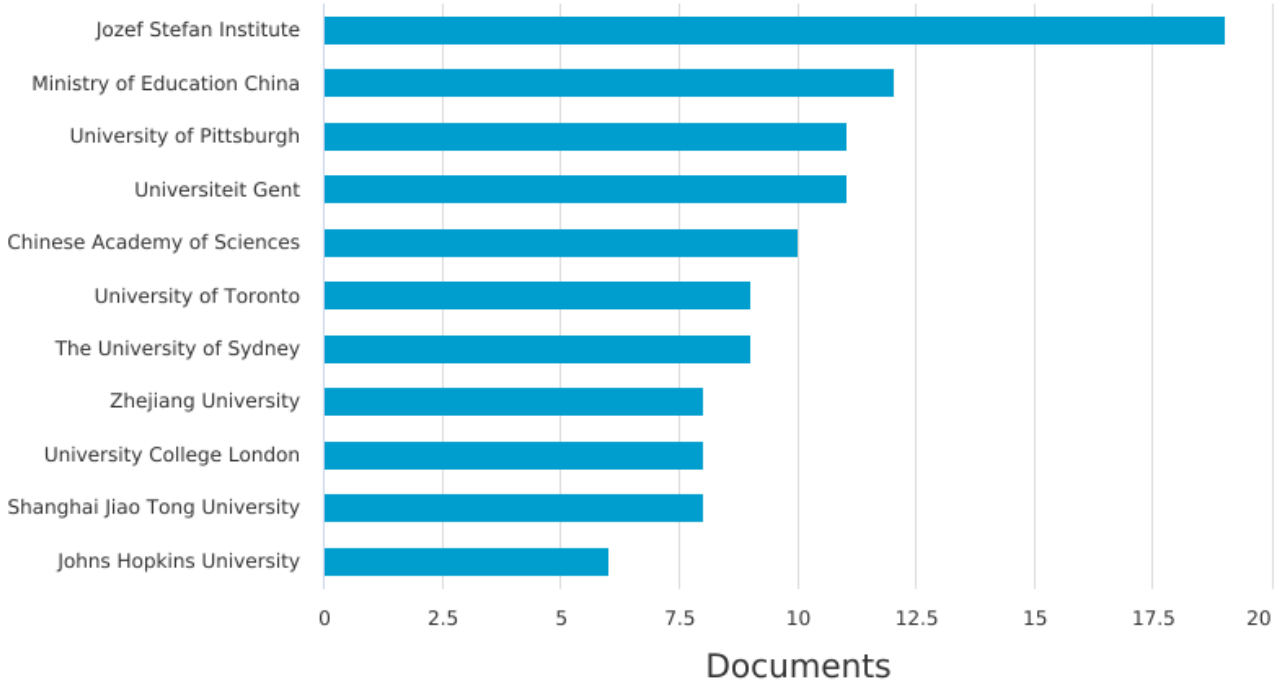

Figure 6: Analysis based on affiliations for publications
Source: http://www.scopus.com (January 2021)

2.6 Analysis based on Sources:

Figure 7 shows the analysis of documents by sources. Most of the research work was published in both “Lecture Notes in Computer Science including Subseries Lecture Notes in Artificial Intelligence” and “Lecture Notes in Bioinformatics and International Journal of Radiation, Oncology Biology Physics”. Additionally, there is less publication in this field.

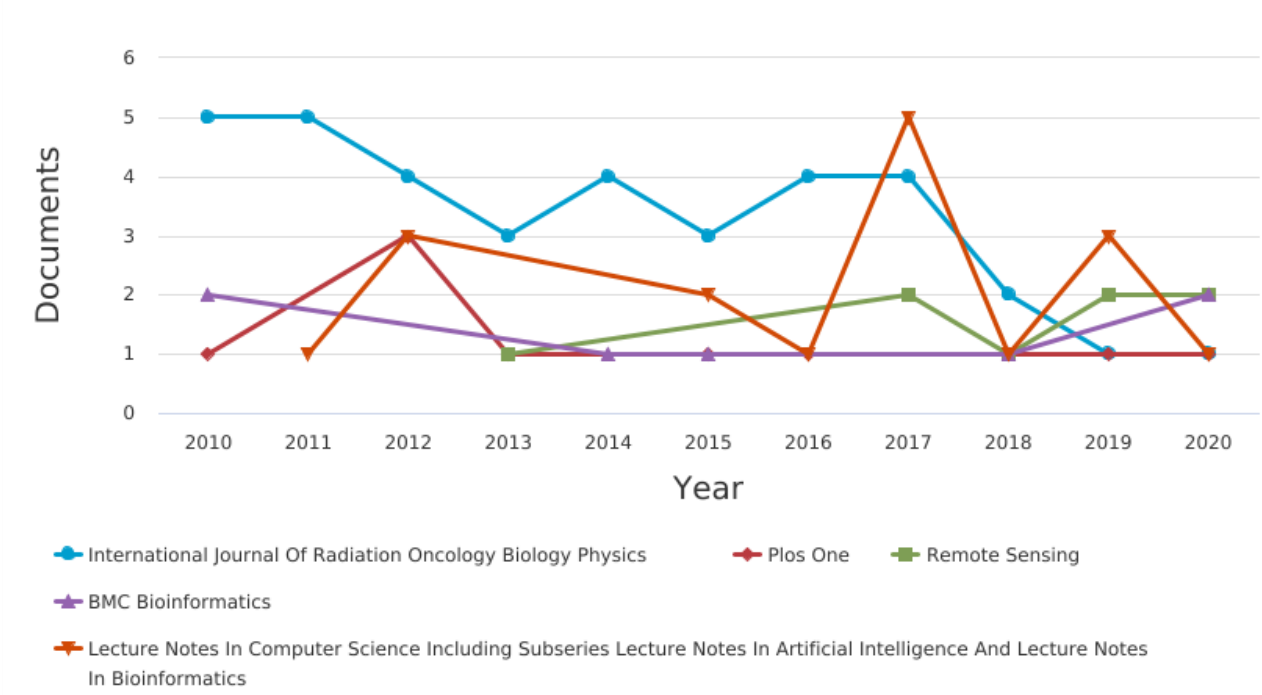

Figure 7: Analysis of documents by sources
Source: http://www.scopus.com (January 2021)

2.7 Analysis based on Funding Sponsors

The National Natural Science Foundation of China has given funded sponsorship for 365 research. The analysis of documents by funding sponsors is shown in Figure 8.

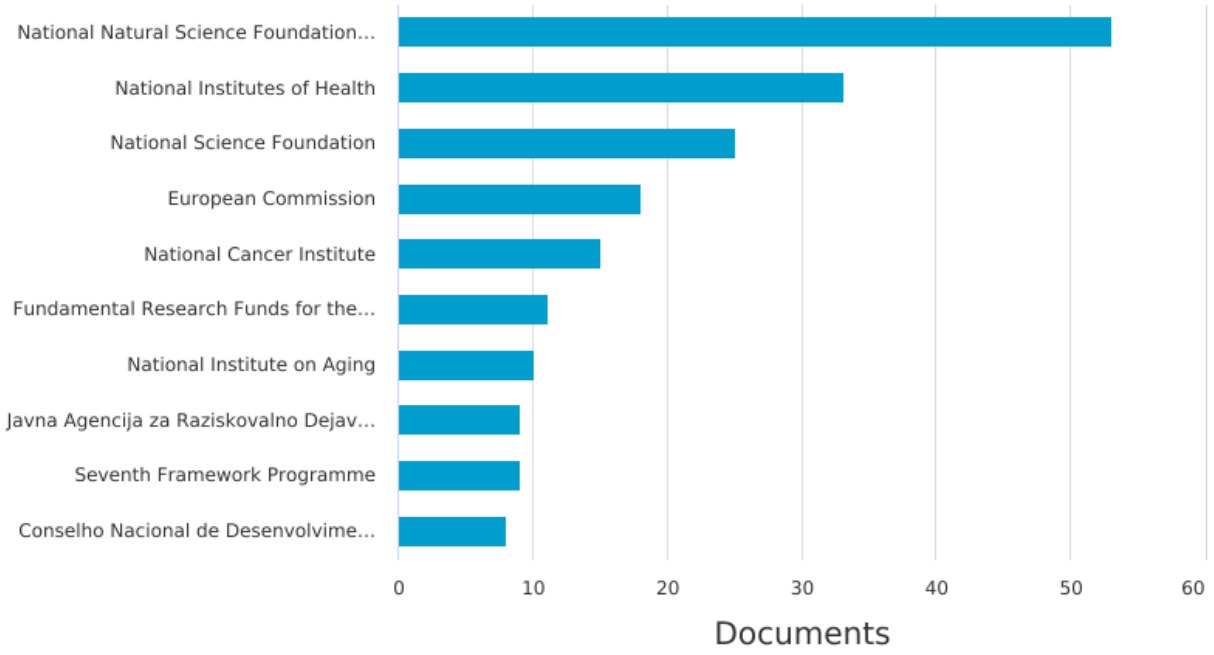

Figure 8: Analysis of documents by funding Sponsor

Source: http://www.scopus.com (January 2021)

2.8 Analysis based on number of publications per author

The authors who contributed to this field are depicted in figure 9. The first twelve authors were considered from the available accessed data from the Scopus database. Author Dzeroski S. contributed 15 documents to this field which was the highest number of contributions.

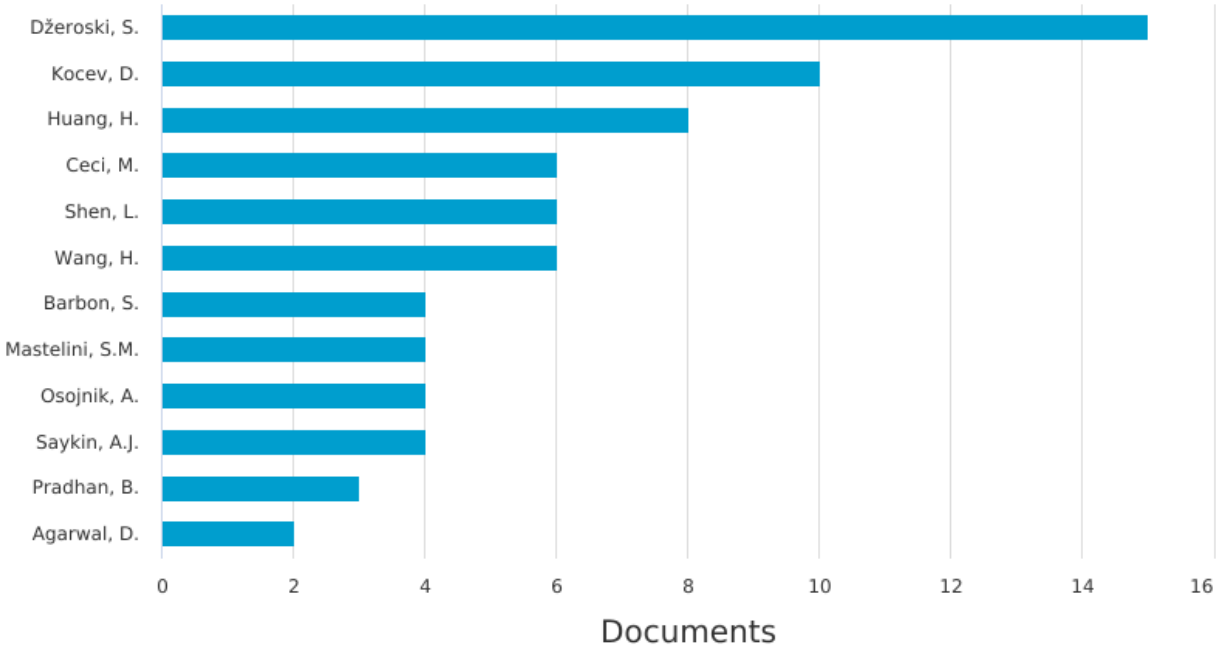

Figure 9: Analysis based on number of publications per author

Source: http://www.scopus.com (January 2020)

2.9 Citation Analysis

Citation analysis is an approach of measuring the influence of an author, an article, a document, or a publication by counting their number of times that particular author, article, document, or publication has been cited by others

Figure 10 shows overall citations for publication per year. In the year 2020, 827 citations were done for the 2756 publications, which were the highest cited to date. Titles of the top-cited papers, with the citations, received to them until the date, till which the data is extracted, are shown in the tabular format. Table 4 shows the detailed information about the research done.

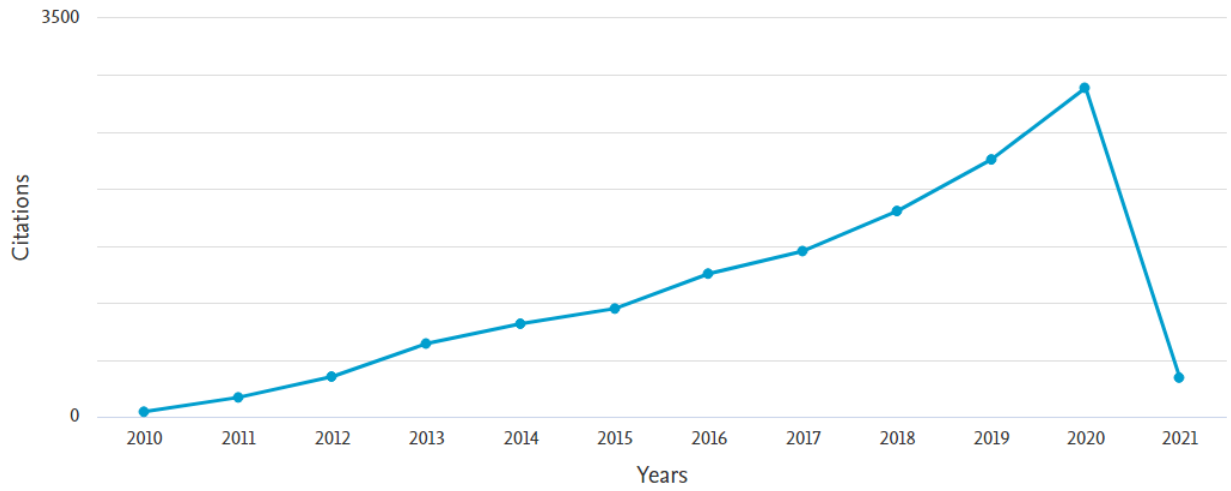

Figure 10: Analysis of papers being cited per year in the area of Predictive Analysis

Source: http://www.scopus.com (January 2021)

Table 4: A citation analysis of top publications in Predictive Analysis

Document Title	<2017	2017	2018	2019	2020	2021	total
Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy	58	40	51	64	56	10	279
Comparison and ranking of different modeling techniques for prediction of site index in Mediterranean mountain forests	94	22	20	32	42	6	216
Multi-target regression via input space expansion: treating targets as inputs	3	12	24	22	36	4	101

Multi-site solar power forecasting using gradient boosted regression trees	0	2	12	30	48	3	95
Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance	35	19	13	14	12	2	95
Accuracy improvements for multi-criteria recommender systems	28	7	32	13	10	2	92
A predictive multiscale computational framework for viscoelastic properties of linear polymers	39	13	6	14	6	0	78
Analysis of daily solar power prediction with data-driven approaches	14	14	13	11	20	3	75
Clustering- and regression-based multi-criteria collaborative filtering with incremental updates	9	10	19	20	15	1	74
Fuzzy regression modeling for tool performance prediction and degradation detection	50	5	2	2	1	0	60
Data-driven multi-touch attribution models	22	8	7	6	9	1	53

Source: http://www.scopus.com (January 2021)

3. NETWORK ANALYSIS

Network analysis is a technique to show the graphical representation of the data matrix. It uses nodes and edges to show their relationship between them. In this representation, relationships are shown using edges and the variables are displayed as nodes of the graph plot. This analysis can be executed using different tools like Cuttlefish, Gephi, Sciencescape, ScoNetV, VOSviewer. By combining different parameters from extracted data of Scopus are used for creating the clusters and visualization are shown in figures 10-16. Different layout algorithms were used along with manual adjustments for the layout while clustering.

Figure 11 shows the network analysis for Cluster of Co-occurrence of Author Keyword. Visualization of the cluster contains 806 items and 48 clusters which represent the co-occurrences of authors keywords. VOSviewer tool is used to show the network visualization of data.

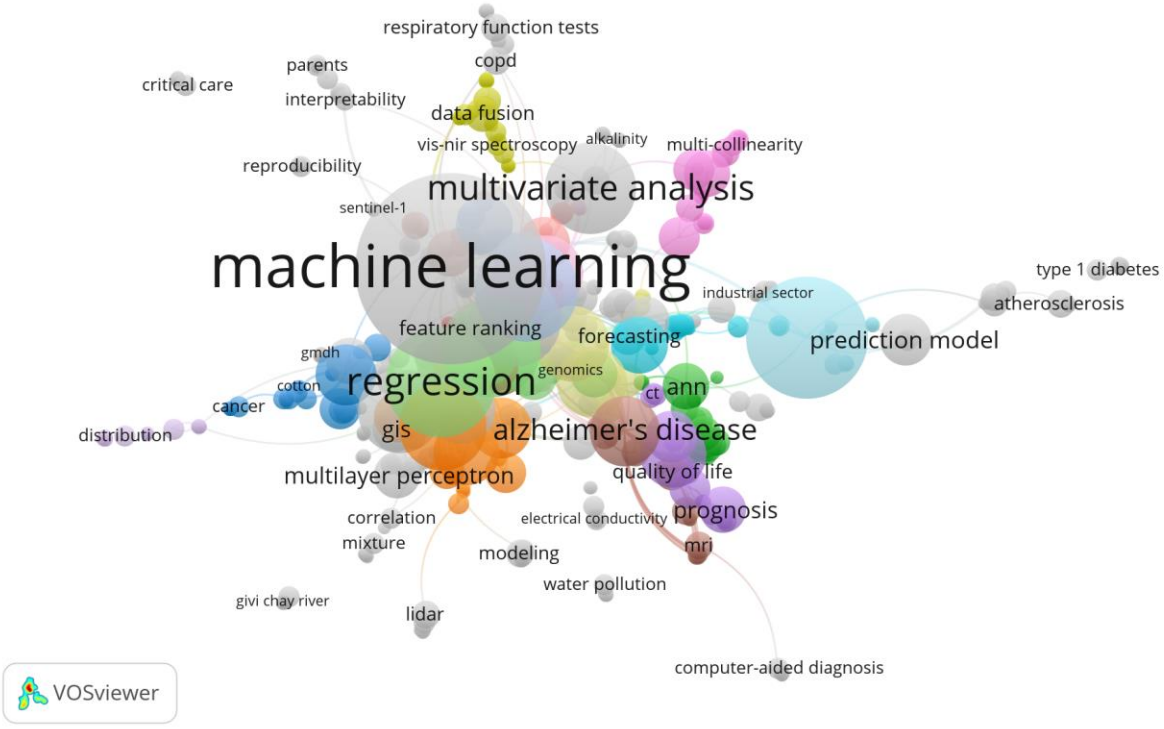

Figure 11: Cluster Visualization of Co-occurrence of Author Keyword
(Source: http://www.scopus.com (January 2021)

Figure 12, shows the visualization as well as the citations received by the documents. The analysis is conducted taking a minimum of two citations per document. The 47 documents were selected with this threshold value 1 and the calculation of the citation link was done accordingly.

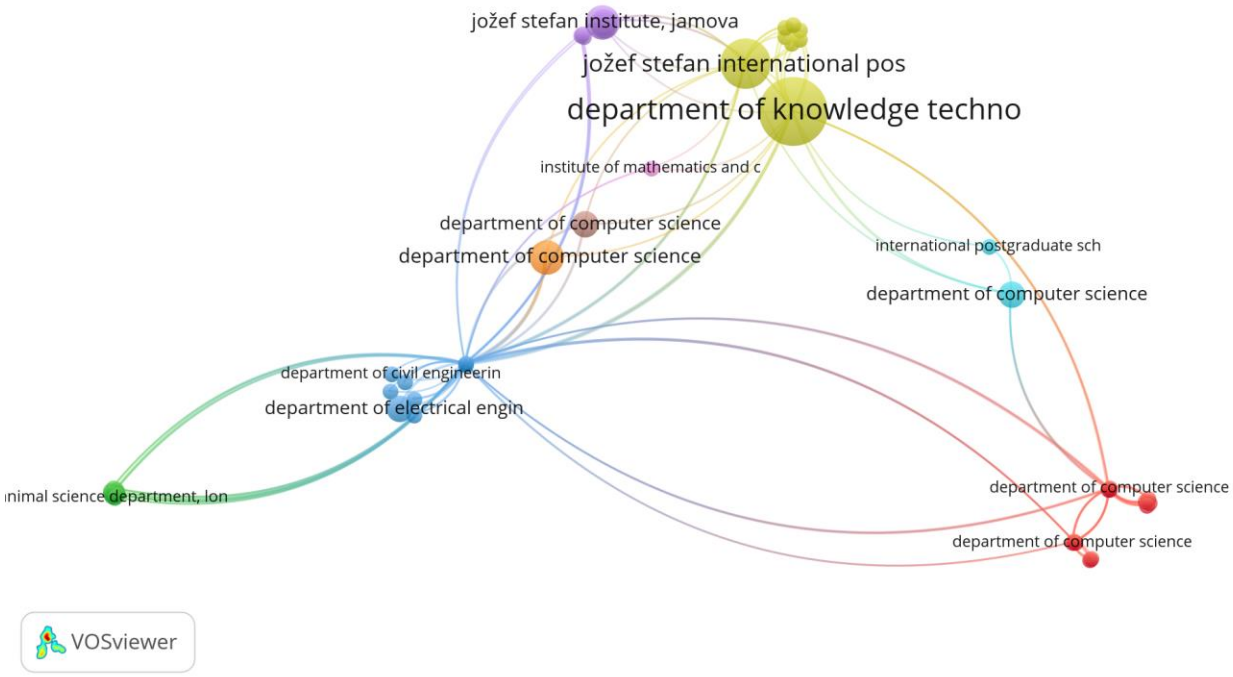

Figure 12: Analysis of Highest Cited Organization
(Source: http://www.scopus.com (January 2021))

Figure 13, shows the visualization based on countries that contributed to the research area. The analysis is conducted taking a minimum of one document per country. The 70 items were selected with this threshold value 1 and the calculation of the cluster was done accordingly. Which form the 13 clusters.

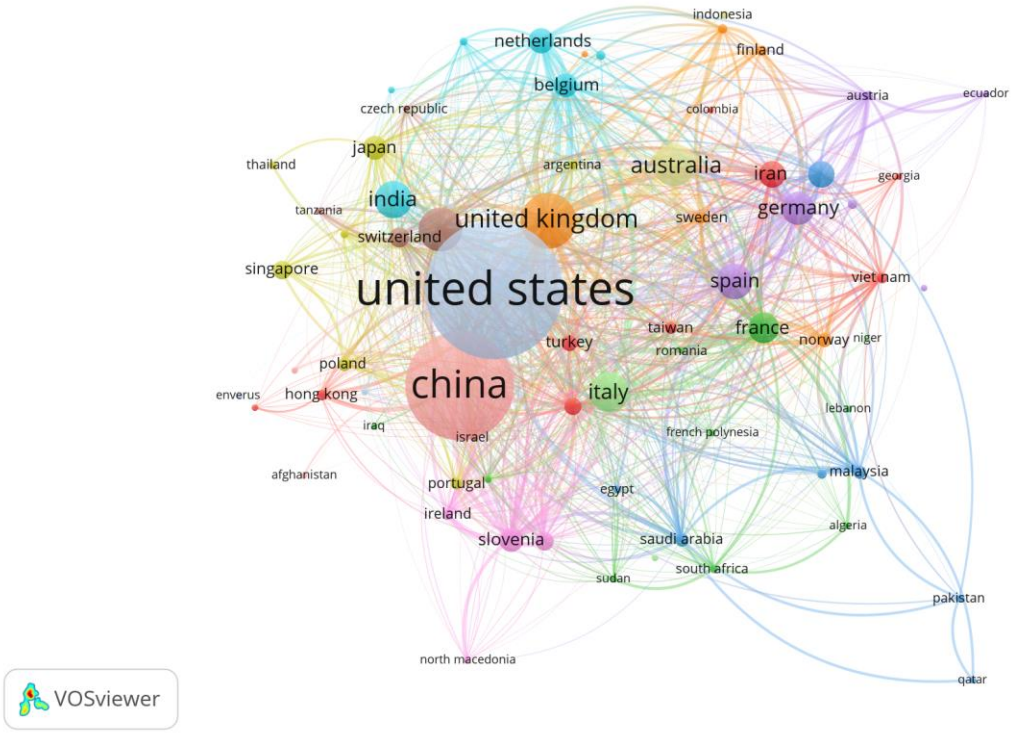

Figure 13: Network Visualization Diagram based on Countries who contributed to this research area. (Source: http://www.scopus.com (January 2021))

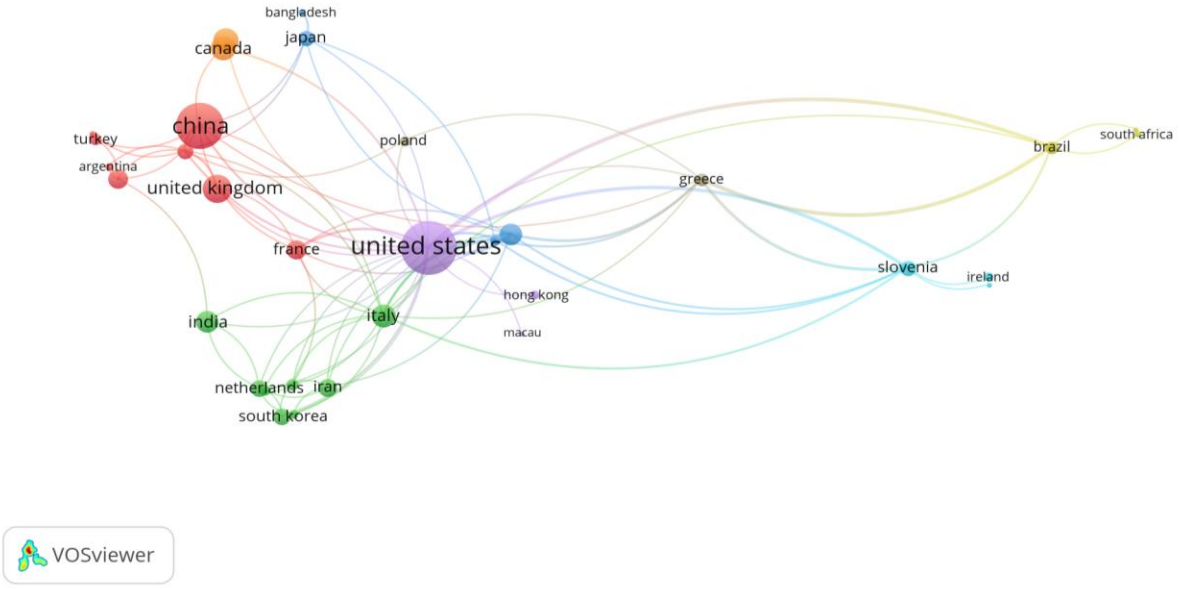

Figure 15: Visualization of Citation of Document based on Countries
(Source: http://www.scopus.com (January 2021))

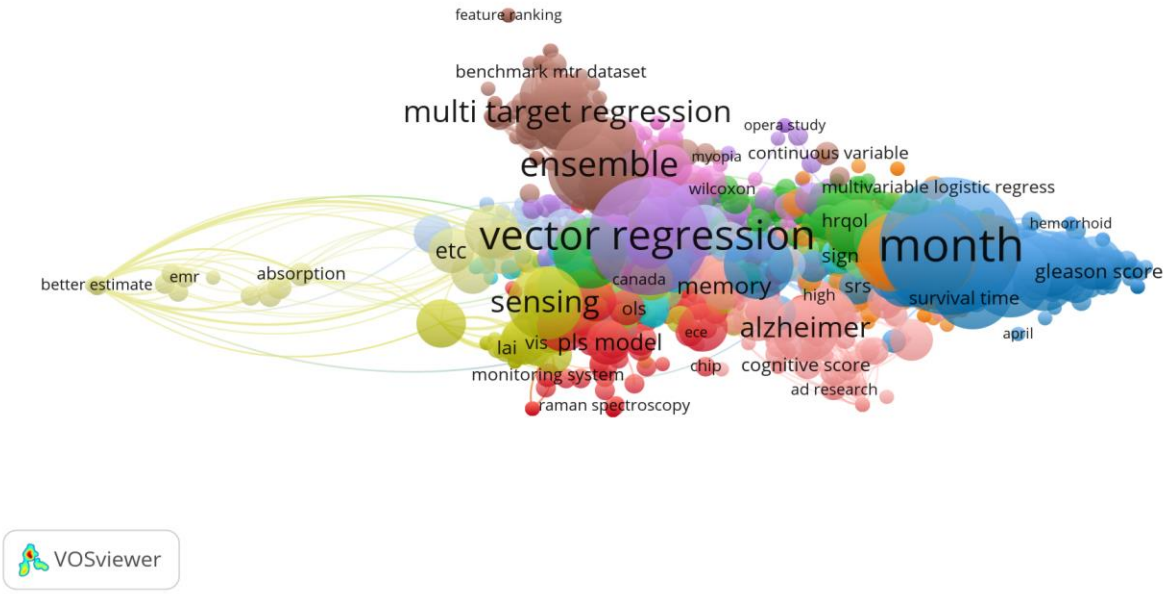

Figure 16 shows the Network of Extracted Data from the Title and Abstract field
(Source: http://www.scopus.com (January 2021))

4. LIMITATIONS OF PRESENT STUDY

In this bibliometric study paper, we used only the Scopus dataset by selecting the publications which are having a combination of keywords. There are different databases like Google Scholar, PubMed, and Web of Science for finding the research documents. Using these datasets we can find the different statistical analyses and visualized data based on the citations of the documents. We used the VOSviwer software tool for the construction and visualization of bibliometric data there are different tools that can be used. We used only the English language and only particular keywords for research publications. One can use different combinations of keywords by adding or removing the same meaning word. In this study paper, we review papers in a certain period from 2010 to 2021 certainly we didn't consider study papers before this period. Therefore, this review paper has the scope for future work to be done.

5. CONCLUSIONS

After analyzing this bibliometric study, we have observed that many researchers are working in the domain of predictive analysis used in different applications from 2006 to date. As predictive analysis is more popular nowadays, most of the publications were in 2020. With the earlier information presented with the help of graphs, it is clear that many journals are publishing their work in this research domain. Different Evolutionary algorithms have been used for prediction. Different clustering techniques are carried out on bibliographic datasets which leads to getting a higher silhouette of data. The analysis based on these algorithms will be focused on further work. The advantage of our study is that we have worked on world data to generate models. These models are relatively easy to implement. However, there are some limitations to the study and these limitations could be extended further as the research goes on in the future.

REFERENCES

[1]Shirish Jeble, "Role of Big Data and Predictive Analytics", International Journal of Automation and Logistics, 2016.

[2]. Fatimetou Zahra Mohamed Mahmoud, "The Application of Predictive Analysis: Benefits, Challenges, and How it can be improved", International Journal of Scientific and Research Publication, 2017.

[3]. Usman Shakeel, Mike Limaco, "Leveraging Cloud-Based Predictive Analytics to Strengthen Audience Engagement", SMPTE Motion Imaging Journal, 2016.

[4]Swati Gupta, "A Regression Modeling Technique on Data Mining", International Journal of Computer Applications, 2015.

[5] Jain, R.K., Smith, K.M., Culligan, P.J., Taylor J.E., “Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy”,2014.

[6] Long, H., Zhang, Z., Su, Y. “Analysis of daily solar power prediction with data-driven approaches”,2014.

[7] Prof. M.S. Prasada Babu, S.Hanumanth,Sastry, “Big Data and Predictive Analytics in ERP Systems for Automating Decision Making Process”,2014.

[8] Galit Shmueli, Otto Koppius, “Predictive vs. Explanatory Modeling in IS Research” , 2001, pp. 10-19

[9] Gurbaxani V, Mendelson H. “Modeling vs forecasting - the case of information- systems spending. IS Research 5(2)”: PP. 180-190,

[10] Dr. CSR Prabhu, “Big Data and Predictive Analytics using data analytics”,2018

[11] Putra, R. E., Izzati, B. M., and F. Dewi, “Point Of Sale (POS) Performance Optimization With Database Synchronization Implementation Using Middleware”. Informatika Mulawarman :Jurnal Ilmiah Ilmu Komputer, 12(2), 123, 2017.

[12] M Mahaputra Hidayat,R Dimas Adityo, Alek Siswanto “Design of Restaurant Billing System (E Bill Resto) by Applying Synchronization of Data Billing in Branch Companies to Main Companies Based on Rest API,2020 International Conference.

[13] Jean-Paul Smets-Solanes, Christophe C´eriny, and Romain Courteaud , “SlapOS: a Multi-purpose Distributed Cloud Operating System Based on an ERP Billing Model”,2011

[14] Melvin Greer, “MedBook: A Cloud-based Healthcare Billing and Record Management System” 2012.

[15] K. Hobson, “What are the Obstacles to Digital Health Records”, 2012.

[16] A.S. Boranbayev, S.N. Boranbayev, “Development and Optimization of Information Systems for Health Insurance Billing”,2010.

[17] Elmroth, E., Marquez, F.G., Henriksson, D. and Ferrera, D.P., “Accounting and billing for federated cloud infrastructures”,2009.

[18] Park, K. W., Han, J., Chung, J. W., & Park, K. H. “Themis: a
mutually verifiable billing system for the cloud computing environment”,2009.

[19]Yujian Zhu, Junming Ma, Bo An, Donggang Cao “Monitoring and Billing of A Lightweight Cloud System Based on Linux Container”, 2017.

[20] Ayyoub, B., Zahran, B., Nisirat, M.A., Al-Taweel, F.M.S., Al Khawaldah, M. “A Proposed cloud-based billers hub using secured e-payments system”,2021.

[21] GINA GUILLAUME-JOSEPH, JAMES S. WASEK, “Improving Software Project Outcomes Through Predictive Analytics”, 2015

[22] Johannes Kunze von Bischhoffshausen, Markus Paatsch, Melanie Reuter, Gerhard Satzger, Hansjoerg Fromm, “An Information System for Sales Team Assignments Utilizing Predictive and Prescriptive Analytics”, 2015

[23] Suraj Pandey, Surya Nepal, Shiping Chen. “A Test-bed for the Evaluation of Business Process Prediction Techniques”, October 2011.
[24] MeryemOuahilal, Mohammed El Mohajir, Mohamed chahhou, BadrEddine El Mohajir. “A Comparative Study of Predictive Algorithms for Business Analytics and Decision Support systems: Finance as a Case Study”. 2016. IEEE
[25] Hina Gulati. “Predictive Analytics Using Data Mining Technique”. 2015. IEEE
[26] Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., &Buyya, R, “Big Data computing and clouds: Trends and future directions.”,2015
[27] Sakineti Subbalakshmi, Dr. CSR Prabhu, “Protagonist of Big Data and Predictive Analytics using data analytics.”,2018
[28] Scopus database: www.scopus.com (Data accessed till 24th January 2021)
[29] VOSviewer download website: https://www.vosviewer.com/download