Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120820

Examining the Trend of Literature on Classification Modelling: A Bibliometric Approach

Hashibah Hamid[1,], Aidi Ahmi[2], Friday Z. Okwonu[3] and Wan Azani Mustafa[4]*

[1]School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[2]Tunku Puteri Intan Safinaz School of Accountancy, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[3]Department of Mathematics, Faculty of Science, Delta State University, P.M.B.1 Abraka, Nigeria
[4]Faculty of Electrical Engineering Technology, Universiti Malaysia Perlis, 02600 UniMAP Arau, Perlis, Malaysia

Received: 11 Feb. 2023, Revised: 27 Jun. 2023, Accepted: 2 Aug. 2023.
Published online: 1 Aug. 2023.

Abstract: This paper analyses and reports various types of published works related to classification or discriminant modelling. This paper adopted a bibliometric analysis based on the data obtained from the Scopus online database on 27th July 2019. Based on the 'keywords' search results, it yielded 2775 valid documents for further analysis. For data visualisation purposes, we employed VOSviewer. This paper reports the results using standard bibliometric indicators, particularly on the growth rate of publications, research productivity, analysis of the authors and citations. The outcomes revealed that there is an increased growth rate of classification literature over the years since 1968. A total of 2473 (89.12%) documents were from journals (n=1439; 51.86%) and conference proceedings (n=1034; 37.26%) contributed as the top publications in this classification topic. Meanwhile, 2578 (92.9%) documents are multi-authored with an average collaboration index of 3.34 authors per article. However, this classification research field found that the famous numbers of authors' collaboration in a document are two (with n=758; 27.32%), three (n=752; 27.10%) and four (n=560; 20.18%) respectively. An analysis by country, China with 1146 (41.30%) published documents thus is ranked first in productivity. With respect to the frequency of citations, Bauer and Kohavi (1999)'s article emerged as the most cited article through 1414 total citations with an average of 70.7 citations per year. Overall, the increasing number of works on classification topics indicates a growing awareness of its importance and specific requirements in this research field.

Keywords: Bibliometric analysis, classification, discrimination, Scopus, visualisation.

1 Introduction

Classification problems are found in theoretical and real-world applications concerning classifying new objects (firms, individuals, plants, etc.) into predetermined groups (agency, class) [1]. The practice of this discrimination is known as supervised classification [2]. Meanwhile, discriminant analysis is the initial method of classification [3]. The main concern of discriminant analysis is to obtain an analytical classification rule, which is able to assign objects accurately to their predefined groups [4,5]. According to [6], in order to predict a group of upcoming objects, discriminant analysis has been broadly used for such purposes.

Classification is a worthwhile exploration field to be studied as it assists and supports decision-making, mainly. Many researchers have examined predictive discriminant analysis to discover classification issues in various real applications. For instance, classification has been employed in medical science to deliver diagnostic evidence such as predicting the condition of a patient [7,8,9,10]. Classification has also been practiced in finance to predict a company's bankruptcy in order to maximize future profits [11,12,13]. Moreover, classification is performable in the area of business marketing to forecast the purchase intention of the consumers in order to investigate the business value of a branded product [4]. According to [14] and [15], classification is found in various areas ranging from education, finance to medicine.

According to [16], classification methods can be divided into three: semi-parametric, parametric and non-parametric methods. Parametric methods are stronger than non-parametric methods as they need less data to produce strong conclusions [17]. Nevertheless, all data points must exhibit a bell-shaped curve that is normally distributed [18]. Examples of parametric methods that are frequently used are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), soft independent modelling of class analogy (SIMCA) [19], partial least squares discriminant analysis (PLSDA) and location model [20,21].

On the other hand, the non-parametric method is distribution-free [22]. Non-parametric methods allow more flexible methods than parametric in accommodating different distributions [23]. Furthermore, it allows for examining and

*Corresponding author e-mail: hashibah@uum.edu.my

introducing data without a prior assumption about the data distribution [24]. For example, no assumption has been made about the distribution of the data for methods like classification and regression trees (CART) as well as for k-nearest neighbor (*k*nn).

Meanwhile, the semi-parametric method is a combination of a non-parametric method and a parametric method [25]. The semi-parametric method estimates the problem regardless of non-smooth measurement functions which consist of both infinite and finite-dimensional unknown parameters and have very weak assumptions [26]. Semi-parametric has an advantage as it does not need any prior knowledge and information to model relation [27]. Logistic discriminant analysis is an example of the semi-parametric method [28].

All these three classification methods have been used extensively to achieve their respective research objectives. Increasingly, the number of publications on classification research is readily available, yet insufficient thought has been given to this study area. In addition, existing scholars' review articles focus primarily on the content, process and methodology of the classification research [29,30,31,32,33,34]. Furthermore, most of the articles adopt classical statistical methods and their hybridisations [35,36,37,38]. To date, there has been a very lack of systematic review of classification research in terms of the attributes and characteristics of its framework.

There is still a lack of comprehensive bibliometric analyses emphasising classification research at the global level that also considers scholarly networks. Therefore, this paper aims to fill this hole by investigating the global trends and scholarly networks involving classification research based on a bibliometric analysis of highly cited articles from 1968 to 2020 taken in the Scopus database. The time of analysis has been decided to capture the progress and the development in classification research since its introduction in 1936 by 39. The analysis was done based on the outputs of publications, journals, citations, authors, author keywords, institutions as well as countries.

2 Methodologies

The analysis of scientific work and publications that relate to classification domains from 1968 to 2020 are presented in this paper. Considering the fact that Scopus is the largest scholarly works database as compared to either Web of Science or PubMed [40,41,42], this paper employed the Scopus database as a basis for extracting prior classification works. The database supplies publication details that include author name, year, source title, document type, subject area, access type, source type, affiliation, country, language and keyword. Further analysis also has been conducted using Microsoft Excel and other bibliometric tools such as Harzing's Publish or Perish and VOSviewer (www.vosviewer.com). To further specify relevant scholarly works on the research domain examined, we explored all publications related to classification works as of the Scopus database using the search terms "classification modelling" OR "classification algorithm" OR "discriminant modelling" OR "discriminant algorithm" in the article title field. This filtering yielded 2,775 total documents for further analysis. The data were retrieved on 27th July 2019.

For the purpose of this paper, a bibliometric approach was conducted using both quantitative and qualitative analysis as well as a mapping of a network of a few bibliometric indicators. The analysis was accomplished based on the data gathered from the Scopus database. Network mapping was performed using VOSviewer to focus on the "network" and "link strength" between co-authorship, author keywords, document titles, and citations. VOSviewer is a free tool for network mapping and visualising bibliographic data [43]. Furthermore, this network analysis is conducted to map the structure and the scope of the discipline, whilst identifying the key research clusters [44,45].

3 Results and Discussion

The analysis of extracted scholarly classification research encompasses document types and source types, annual growth, language of the document, subject area, keywords analysis, country productivity, authorship and citation analysis. The results of this paper are mostly presented in terms of frequency, percentage, graph and visualisation map.

3.1 Document and Source Types

The data is first analysed based on its document and source types using descriptive statistics, i.e., frequency and percentage. Summarising depicted in Table 1 shows that the documents published on classification spread into 10 document types. The result further indicates that only two types of documents successfully attracted scholars to publish their research outputs, i.e., articles and conference papers. Almost half of the entire published documents are in the form of articles (49.84%), followed by conference papers with 47.64%. Other types of documents have recorded less than 1% of the total published document. There is 0.47%, which we do not know the type of document published by the researchers.

Table 1: Documents type

Document Type	**No. of Document (*n*)**	**Percentage (%)**
Article	1383	49.84
Conference Paper	1322	47.64
Book Chapter	23	0.83
Review	21	0.76
Erratum	5	0.18
Letter	4	0.14
Book	2	0.07
Conference Review	1	0.04
Editorial	1	0.04
Undefined	13	0.47
Total	**2775**	**100.00**

Figure 1 shows the documents that were grouped into six different source types. Journal represents the uppermost type of source (1,439; 51.86%) followed by the conference proceedings (1,034; 37.26%). Book series contribute 10.49% (291 documents) to the total number of publications.

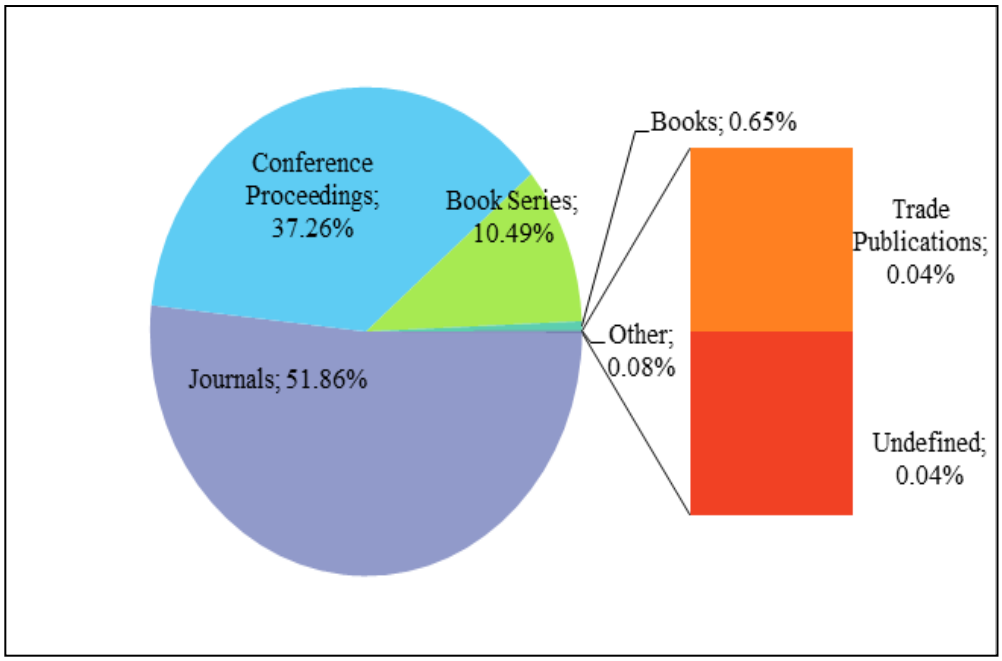

Fig. 1: The type of source where the documents are published

3.2 Evolution of Published Studies by Year

Table 2 summarizes the frequency and percentage of publication by year on classification works from 1968 to 2020 taken from a Scopus database. The first published research related to the domain examined is in 1968, with only three documents (0.11%). The evolution of the related publication is somewhat very slow in the next few years until it starts picking up in 1994 with an average of 13 documents (0.47%). The highest numbers of publications are observed in 2016, 2017 and 2018 with 226 (8.14%), 240 (8.65%) and 276 (9.95%) documents respectively. However, it did decline slightly in 2019 as only 174 documents (6.27%) are published. This is because the data we retrieved from Scopus is only up to July, so its result is not appropriate to compare here. As for 2020, there are already three documents with 0.11% published.

Table 2: Year of publications

Year	**No. of Document (*n*)**	**Percentage (%)**	**Year**	**No. of Document (*n*)**	**Percentage (%)**
1968	3	0.11	1996	15	0.54
1969	1	0.04	1997	10	0.36
1973	1	0.04	1998	16	0.58

1974	2	0.07	1999	11	0.40
1975	1	0.04	2000	16	0.58
1976	2	0.07	2001	27	0.97
1977	2	0.07	2002	33	1.19
1978	2	0.07	2003	28	1.01
1979	4	0.14	2004	58	2.09
1980	3	0.11	2005	68	2.45
1981	5	0.18	2006	75	2.70
1982	5	0.18	2007	103	3.71
1983	4	0.14	2008	106	3.82
1984	6	0.22	2009	132	4.76
1985	6	0.22	2010	140	5.05
1986	5	0.18	2011	157	5.66
1987	8	0.29	2012	180	6.49
1988	6	0.22	2013	181	6.52
1989	10	0.36	2014	191	6.88
1990	2	0.07	2015	190	6.85
1991	5	0.18	2016	226	8.14
1992	3	0.11	2017	240	8.65
1993	7	0.25	2018	276	9.95
1994	13	0.47	2019	174	6.27
1995	13	0.47	2020	3	0.11
Total	**119**	**4.29**	**Total**	**2656**	**95.71**

Figure 2 further displays the percentage of publications on the classification domain based on years from 1968 to 2020. It demonstrates that the volume of publications increases from year to year, but it slightly dropped in 2019 as we can observe that the highest percentages of publications were from 2016 to 2018, and peaked in 2018 at 9.95%.

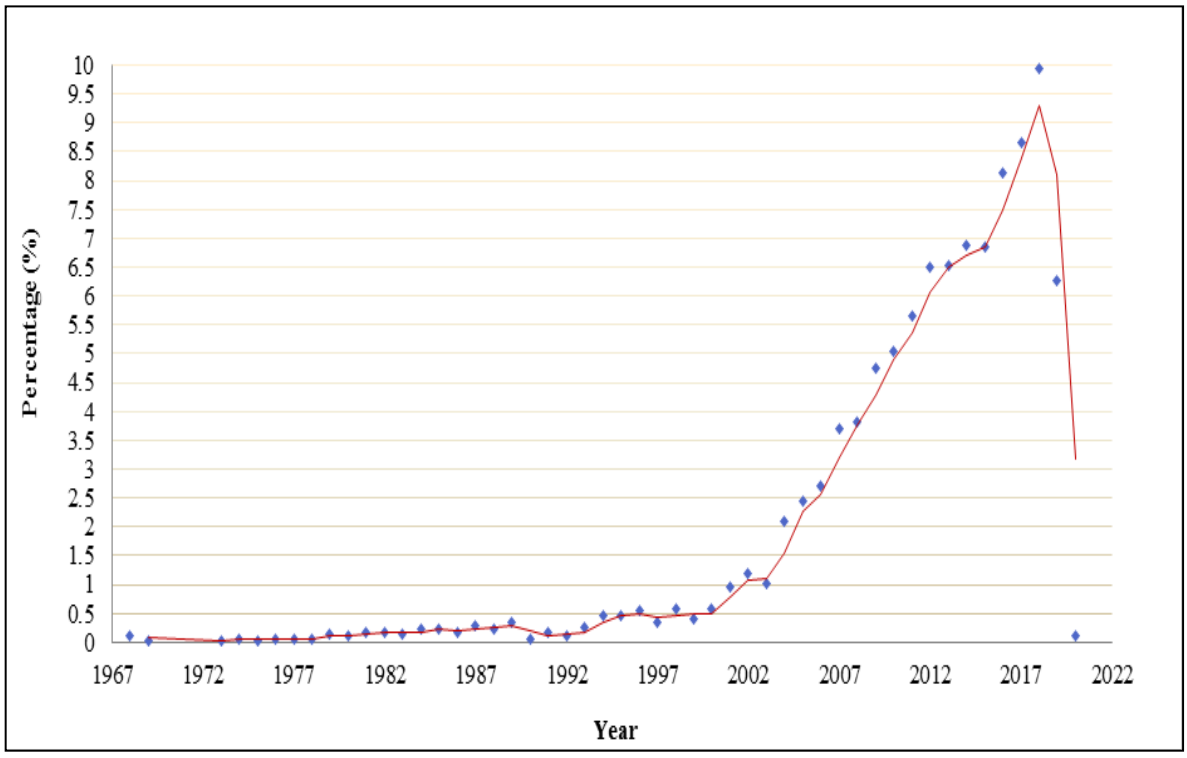

Fig. 2: Percentage of publications on classification topics from 1968 to 2020

3.3 Languages of Documents

Table 3 expresses that English is the common language for the article obtained from the Scopus database, which gives the highest percentage of its implementation with 86.06%, as equivalent to 2,390 documents. The following highest commonly encountered language is Chinese, with 334 documents contributing to 12.03%. There are eight other languages that have been used in the documents, including Turkish, Russian, Spanish, Portuguese, Japanese, Korean, German and Slovenian with their frequencies and percentages as recorded in Table 3, respectively. We notified that there are two documents published in dual languages.

Table 3: Languages used for publications

Languages	**Frequency (*n*)**	**Percentage (%)**
English	2390	86.06
Chinese	334	12.03
Turkish	11	0.40
Russian	8	0.29
Spanish	6	0.22
Portuguese	4	0.14
Japanese	3	0.11
Korean	3	0.11
German	2	0.07
Slovenian	1	0.04
Undefined	15	0.54
Total	**2777**	**100.00**

3.4 Subject Area

This study then discusses the published documents based on the subject area as summarised in Table 4. Overall, the distribution indicates that research on classification emerges in various subject areas ranging from technology, engineering, mathematics, science, healthcare, social science and many more. As reported, more than one-third of the documents are in the computer science area (34.38%), followed by engineering and mathematics with 23.11% and 10.92% respectively. For other areas, it presents less than 5% of publications. The top three percentages for subject areas, i.e., computer science, engineering and mathematics, where the research published related to the classification domain also can be spotted in Figure 3.

Table 4: Subject area

Subject Area	**No. of Documents (*n*)**	**Percentage (%)**
Computer Science	1647	34.38
Engineering	1107	23.11
Mathematics	523	10.92
Physics and Astronomy	228	4.76
Earth and Planetary Sciences	197	4.11
Medicine	154	3.21
Materials Science	116	2.42
Social Sciences	105	2.19
Biochemistry, Genetics and Molecular Biology	98	2.05
Decision Sciences	90	1.88
Environmental Science	67	1.40
Chemistry	63	1.31
Business, Management and Accounting	62	1.29
Agricultural and Biological Sciences	56	1.17

Neuroscience	55	1.15
Energy	49	1.02
Multidisciplinary	41	0.86
Chemical Engineering	37	0.77
Health Professions	20	0.42
Pharmacology, Toxicology and Pharmaceutics	20	0.42
Economics, Econometrics and Finance	14	0.29
Psychology	13	0.27
Arts and Humanities	12	0.25
Immunology and Microbiology	11	0.23
Nursing	2	0.04
Undefined	4	0.08
Total	**4791**	**100.0**

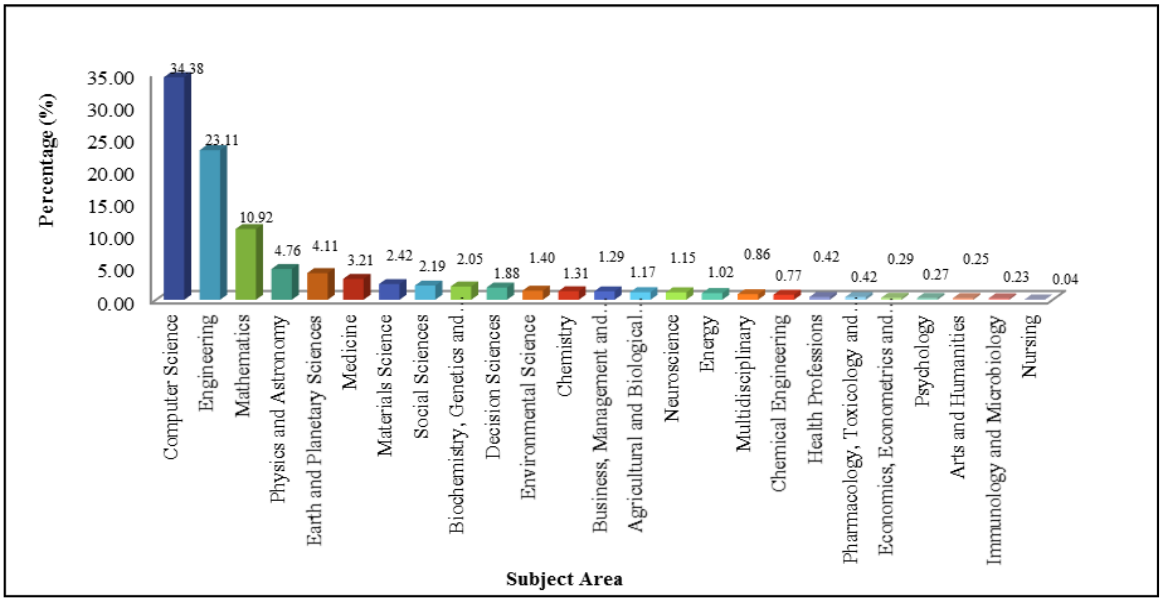

Fig. 3: Documents by subject area

3.5 Source Title

Next, this paper debates the top 20 source titles published by the authors. Table 5 demonstrates that "Lecture Notes in Computer Science including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics" is the primary choice of the source title (with 4.5%; *n*=125) by researchers in publishing their documents. It is followed by the source entitled "Proceedings of SPIE the International Society for Optical Engineering" (2.05%; *n*=57), "Advances in Intelligent Systems and Computing" (1.26%; *n*=35) and "International Geoscience and Remote Sensing Symposium IGARSS" (1.23%; *n*=34). The other source titles of publications on classification topics are shown in Table 5 with their respective percentages (all show less than 1.0%).

Table 5: Top 20 source titles

Source Title	**No. of Document (*n*)**	**%**
Lecture Notes in Computer Science including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics	125	4.50
Proceedings of SPIE The International Society for Optical Engineering	57	2.05
Advances in Intelligent Systems and Computing	35	1.26
International Geoscience and Remote Sensing Symposium IGARSS	34	1.23

Applied Mechanics and Materials	27	0.97
Communications in Computer and Information Science	25	0.90
Lecture Notes in Electrical Engineering	22	0.79
Moshi Shibie Yu Rengong Zhineng Pattern Recognition and Artificial Intelligence	21	0.76
ACM International Conference Proceeding Series	18	0.65
Advanced Materials Research	17	0.61
Tien Tzu Hsueh Pao Acta Electronica Sinica	16	0.58
Dianzi Yu Xinxi Xuebao Journal of Electronics and Information Technology	15	0.54
International Journal of Applied Engineering Research	15	0.54
International Journal of Remote Sensing	15	0.54
Indian Journal of Science and Technology	14	0.50
Procedia Computer Science	13	0.47

3.6 Keywords Analysis

Using VOSviewer, based on ten minimum numbers of occurrences, the author keywords were mapped (see Figure 4). The figure indicates the strength of the association among those keywords. Any keywords that have similar colour are commonly listed together. As an example, the figure implies that classification, support vector machine, fault diagnosis, document classification, data classification, multi-class classification, ant colony optimisation and ensemble learning are closely related and typically co-occur together.

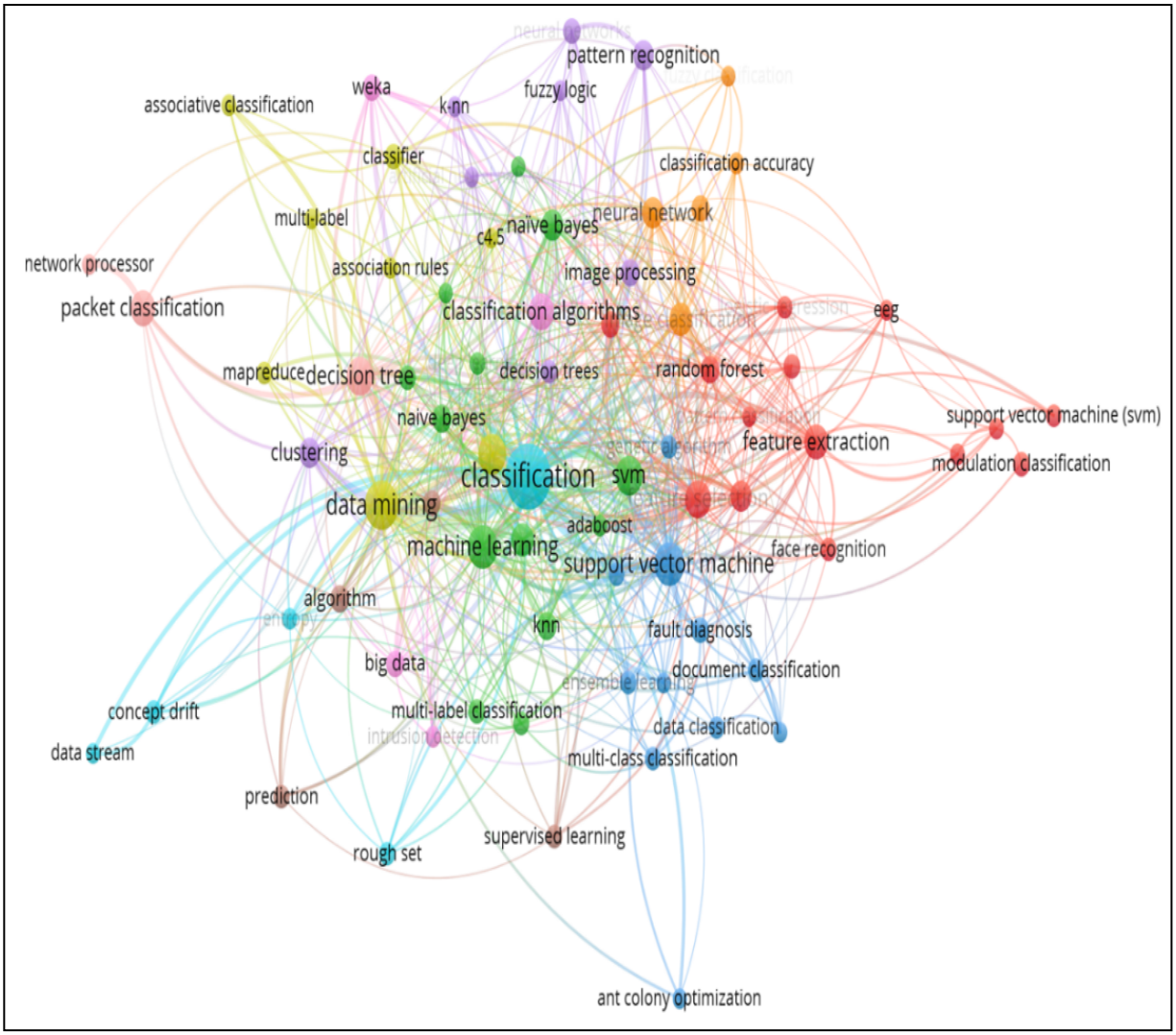

Fig. 4: Network visualisation map of the authors' keywords with ten minimum numbers of occurrences

A different result is obtained when the authors' keywords were visualised with 20 minimum numbers of occurrences, as shown in Figure 5. As displayed, support vector machine, image classification, feature extraction, pattern recognition, remote sensing and neural network have the same colour indicating that these keywords are closely related and usually co-occur together [46].

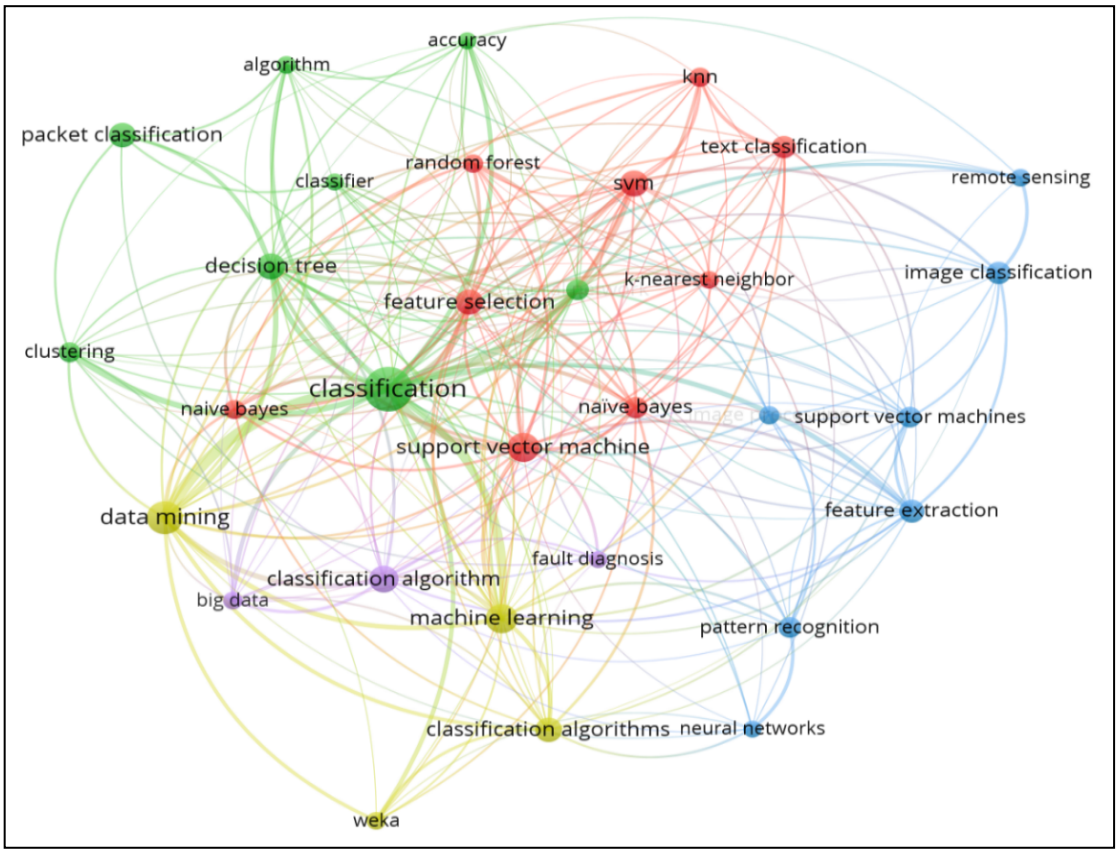

Fig. 5: Network visualisation map of the authors' keywords with 20 minimum numbers of occurrences

In the meantime, this paper also investigates the keywords that appear more than 100 times, as revealed in Table 6. Algorithm(s), Classification Algorithm(s) and Classification (of Information) are the three keywords with the highest occurrences (then, data mining and classification) for the research domain examined with percentages of 50.99%, 42.96% and 32.61%, respectively. Some other keywords show lower than 20% and many of them are under 10% of appearances as can be seen in Table 6.

Table 6: Keywords appear more than 100 times

Author Keywords	**No. of Document (*n*)**	**Percentage (%)**
Algorithm / Algorithms	1415	50.99
Classification Algorithm / Classification Algorithms	1192	42.96
Classification (of Information)	905	32.61
Data Mining	517	18.63
Classification	511	18.41
Support Vector Machine / Support Vector Machines	447	16.11
Learning Systems	256	9.23
Decision Trees	240	8.65
Artificial Intelligence	219	7.89
Learning Algorithms	202	7.28
Feature Extraction	192	6.92
Article	189	6.81
Image Classification	189	6.81
Classification Accuracy	174	6.27

Neural Networks	173	6.23
Human	158	5.69
Pattern Recognition	149	5.37
Machine Learning	140	5.05
Remote Sensing	127	4.58
Text Processing	119	4.29
Trees (mathematics)	118	4.25
Humans	114	4.11
Clustering Algorithms	111	4.00
Image Processing	106	3.82
Classifiers	102	3.68
Signal Processing	100	3.60

3.7 Most Influential Countries

Table 7 lists the top 20 counties that contributed to the publications on classification topics since 1968. The highest one is 41.30% (n=1146) from China, followed by the United States (13.05%; n=362) and India (9.12%; n=253) are among the most productive worldwide in this research area. The other listed countries also contributed to the publication outputs with their respective percentages (see Table 7).

Table 7: Countries contributed to the publications

Country	**No. of Document (*n*)**	**Percentage (%)**
China	1146	41.30
United States	362	13.05
India	253	9.12
United Kingdom	79	2.85
South Korea	67	2.41
Turkey	65	2.34
Germany	64	2.31
France	60	2.16
Canada	57	2.05
Italy	54	1.95
Taiwan	53	1.91
Malaysia	50	1.80
Japan	48	1.73
Spain	47	1.69
Iran	45	1.62
Russian Federation	44	1.59
Australia	38	1.37
Brazil	37	1.33
Netherlands	26	0.94
Greece	24	0.86

3.8 Most Influential Institutions

There are 10 Institutions of interest in which most of the outputs published more than 20 documents in the classification area, as shown in Table 8. Chinese Academy of Sciences, Ministry of Education China and Tsinghua University are among the top three Institutions that contributed more than 20 documents to the publications in the examined domain from 1968 to 2020. Meanwhile, other Institutions also gained publications as depicted in the mentioned table.

Table 8: Institution contributed to the publications with more than 20 documents

Name of Institution	**No. of Document (*n*)**	**Percentage (%)**
Chinese Academy of Sciences	54	1.95
Ministry of Education China	47	1.69
Tsinghua University	37	1.33
University of Electronic Science and Technology of China	29	1.05
Xidian University	26	0.94

Harbin Institute of Technology	24	0.86
Beijing University of Posts and Telecommunications	23	0.83
Wuhan University	22	0.79
Harbin Engineering University	21	0.76
Nanjing University of Science and Technology	20	0.72

3.9 Authorship Analysis

The output of publication can be measured by the number of authors that contribute to the success of the work of a document. It can be measured as a single-author or multi-authored publication. As shown in Table 9, a total of 197 (7.10%) documents were written by single-authored whilst the rest of the documents were produced by multi-authored. The most glamorous number of authors in the documents for this research area is ranging from two to four, with a mean of 690 documents carrying an average of 24.87%. However, here we have one document with 22 authors, and surprisingly there are also 75 authors in a document for this classification area, each carrying 0.04%.

Table 9: Number of author(s) per document

Author Count	**No. of Document (*n*)**	**Percentage (%)**
1	197	7.10
2	758	27.32
3	752	27.10
4	560	20.18
5	261	9.41
6	125	4.50
7	54	1.95
8	25	0.90
9	12	0.43
10	15	0.54
11	4	0.14
12	1	0.04
13	3	0.11
14	2	0.07
15	1	0.04
16	1	0.04
19	2	0.07
22	1	0.04
75	1	0.04
Total	**2775**	**100.00**

The most productive classification authors within the investigation period are listed in Table 10 (only authors with more than five documents are listed here). The top authors are Sun X. with ten classification-related documents (0.36%), whilst Hu Z. P. and Wang Z. each with nine documents (0.32%) as retrieved from the Scopus database.

Table 10: Most influential authors (more than 5 documents)

Author's Name	**No. of Document (*n*)**	**Percentage (%)**
Sun, X.	10	0.36
Hu, Z.P.	9	0.32
Wang, Z.	9	0.32
Freitas, A.A.	8	0.29
Brazdil, P.	7	0.25
Jiao, L.	7	0.25
Otero, F.E.B.	7	0.25
Liu, R.	6	0.22
Carugati, I.	5	0.18
Delimata, P.	5	0.18
Feng, X.	5	0.18
Hu, X.	5	0.18
Kawata, Y.	5	0.18

Li, J.	5	0.18
Liu, H.	5	0.18
Moriyama, N.	5	0.18
Niki, N.	5	0.18
Ohmatsu, H.	5	0.18
Shi, Y.	5	0.18
Skowron, A.	5	0.18
Suraj, Z.	5	0.18
Wang, G.	5	0.18

This study identified eight distinct clusters of collaboration networks based on the co-authorship, (see Figure 6). The figure shows the high level of connection of the collaborative networks for classification research. The analysis revealed that Wang J., Wang Z., Liu Z., Zhang X. and Wang X were among the top productive authors as they appeared with a big circle for each colour in the diagram. This finding also indicates how influential their works are in classification research and the strong interests of these authors in this field [46].

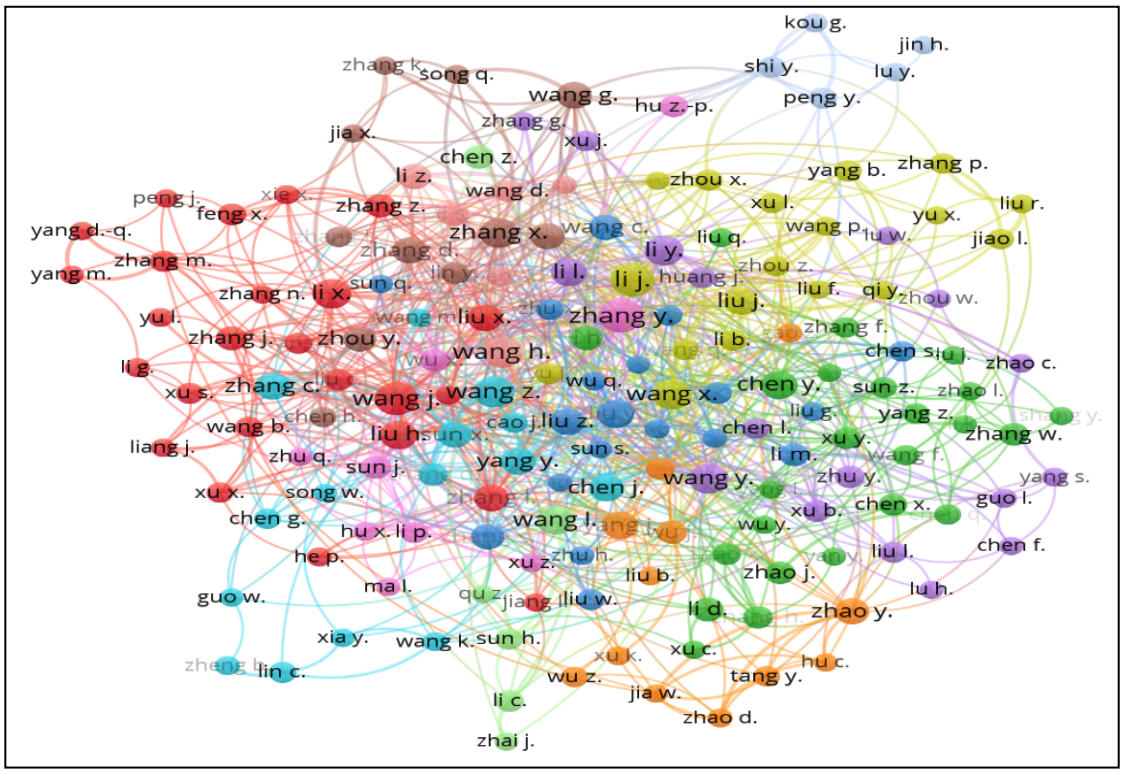

Fig. 6: Network visualisation map of the co-authorship
Note: Unit of analysis = Authors; Counting method = Fractional counting; Minimum number of documents of an author = 5; Minimum number of citations of an author = 5

A different network visualisation map of co-authors is obtained (see Figure 7) when the indicators based on the minimum number of documents of an author, and the minimum number of citations of an author was set to 10. This diagram clearly shows the network for the most active and highly influenced authors in this research field.

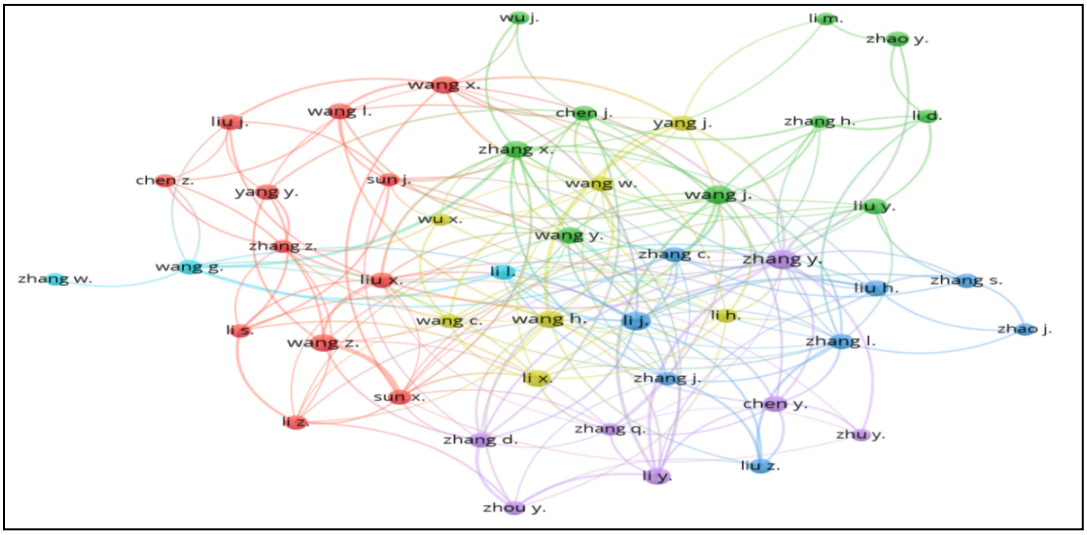

Fig. 7: Network visualisation map of the co-authorship
Note: Unit of analysis = Authors; Counting method = Fractional counting; Minimum number of documents of an author = 10; Minimum number of citations of an author = 10

Meanwhile, the co-authorship network based on the country as a unit of analysis has been generated and analysed (see Figure 8). Figure 8 exposes the collaboration networks among the most productive countries. China was extremely linked to the United States, India, United Kingdom and South Korea. This result is consistent with the outcome in Table 7. The overall strength of the relationship between these four countries and China was 68.73%. This finding indicates that those four countries are participating in a significant proportion of China's network. The findings have identified the following four clusters: countries surrounding China (light purple cluster), countries surrounding the United States (light brown cluster), countries surrounding India (light blue cluster), and countries surrounding the United Kingdom (dark blue cluster). Notably, Malaysia was also listed as a productive country in contributing the documents in the area of classification, and the co-authorship has collaborated with countries such as Indonesia, Taiwan and Turkey.

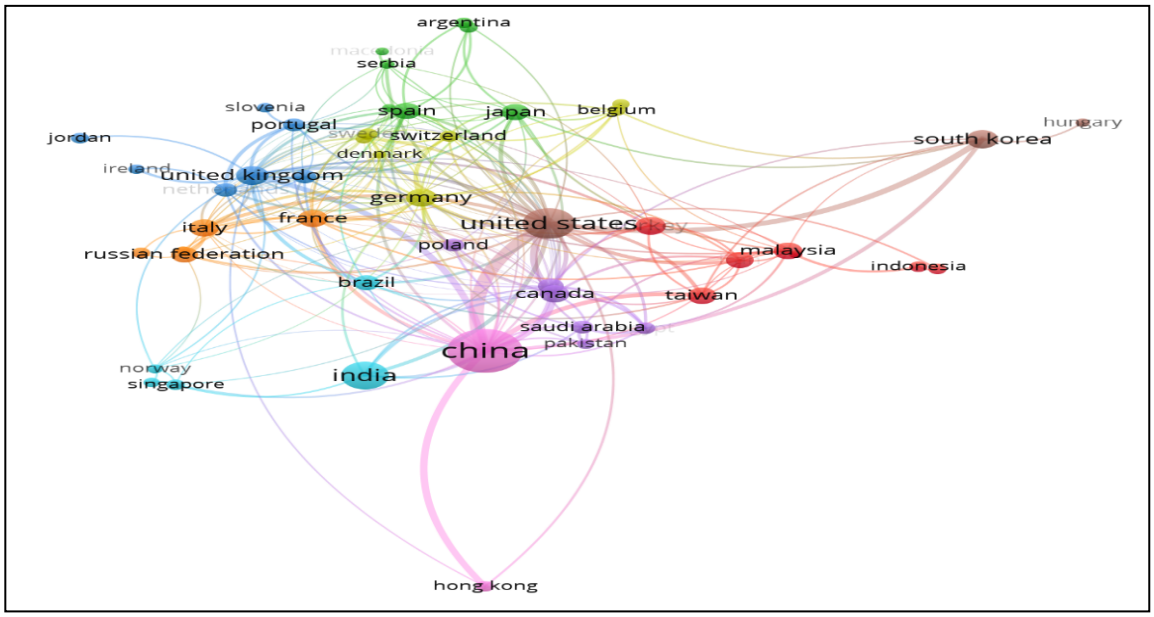

Fig. 8: Network visualisation map of the co-authorship
Note: Unit of analysis = Countries; Counting method = Fractional counting; Minimum number of documents of a country = 5; Minimum number of citations of a country = 5

3.10 Text Analysis

The term co-occurrence based on the title can imitate the research hotspots in a particular domain of research, providing supplementary support for systematic and scientific studies [47]. The term co-occurrence network of classification literature was produced using VOSviewer (see Figure 9). The figure shows that the nodes and word size represent the weightiness of the nodes. If the size of the node and word is bigger, then the weight is larger [48]. The distance and line size between the two nodes also discloses the strength of the connection between them. A shorter distance usually exposes a stronger connection [48]. If the size of the line is thicker, then, there will be more co-occurrence they have [49]. The nodes that share the same colour are grouped into one cluster. VOSviewer has generated the co-occurrence of terms based on the title into five clusters. The term "classification algorithm" has the biggest node and word. It implies that this term has the highest frequency of co-occurrence in the title based on classification-related publications. It is followed by other terms such as image classification algorithm, research, model, neural network, performance analysis and many more as can be seen in Figure 9.

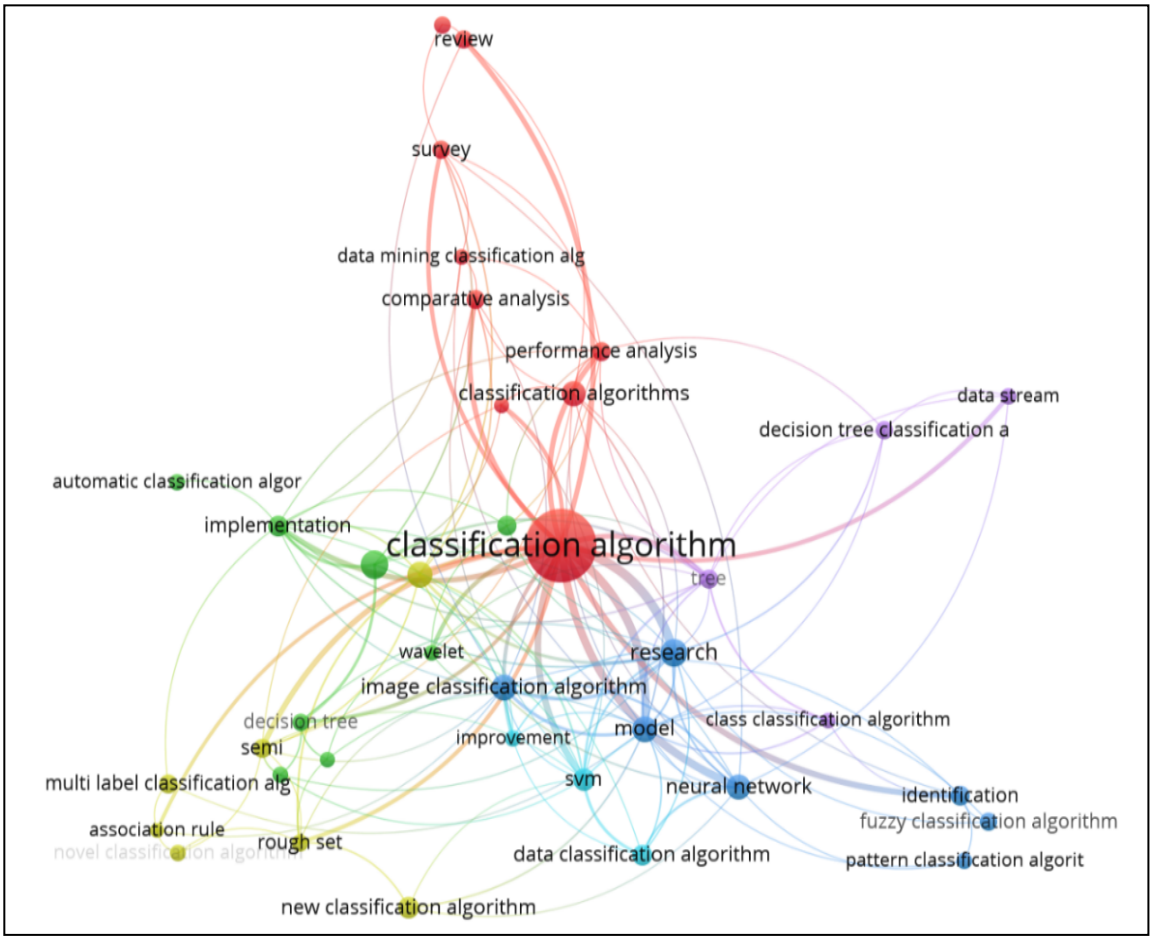

Fig. 9: VOSviewer visualisation of a term co-occurrence network based on title fields (Binary Counting)

3.11 Citation Analysis

We report citation analysis as citation metrics and disclosed the 20 most cited articles in the classification. Table 11 summarizes the citation metrics for the extracted documents on 27th July 2019. Table 11 shows the volume of citations with average citations per year for all extracted documents. As indicated, approximately 20,803 citations were reported within this 50-year period for 2,775 extracted articles with an average of 407.9 citations each year.

Table 11: Citations metrics

Metrics	**Data**
Publication Years	1968-2020
Citation Years	51 (1968-2019)

Papers	2775
Citations	20803
Cites/Year	407.90
Cites/Paper	7.50
Authors/Paper	3.34
h-index	64
g-index	120

Meanwhile, Table 12 reveals the 20 most influential documents as they are the top 20 based on how many times they are being cited as reported by Scopus. The document entitled "Empirical Comparison of voting classification algorithms: bagging, boosting, and variants" by [50] that was published in 1999 has obtained the highest number of citations with a total citation of 1,414 or an average of 70.7 per year. Meanwhile, Lotte et al. discuss the topic "A review of classification algorithms for EEG-based brain-computer interfaces" published in 2007 in "Journal of Neural Engineering". Its total number of citations was the second highest at 1,409 with an average of 117.4 citations per year. Other examples of the top 20 most influential documents can be seen in Table 12.

Table 12: Most influential papers (Top 20)

Authors	**Year**	**Title**	**Source Title**	**Total Citations**	**Citations per Year**
Bauer E., Kohavi R.	1999	Empirical comparison of voting classification algorithms: bagging, boosting, and variants	Machine Learning	1414	70.7
Lotte F., Congedo M., Lécuyer A., Lamarche F., Arnaldi B.	2007	A review of classification algorithms for EEG-based brain-computer interfaces	Journal of Neural Engineering	1409	117.4
Lim T.-S., Loh W.-Y., Shih Y.-S.	2000	Comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms	Machine Learning	668	35.2
Posner K., Oquendo M.A., Gould M., Stanley B., Davies M.	2007	Columbia Classification Algorithm of Suicide Assessment (C-CASA): Classification of suicidal events in the FDA's pediatric suicidal risk analysis of antidepressants	American Journal of Psychiatry	518	43.2
Aggarwal C.C., Zhai C.	2012	A survey of text classification algorithms	Mining Text Data	478	68.3
Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., Vanthienen J.	2003	Benchmarking state-of-the-art classification algorithms for credit scoring	Journal of the Operational Research Society	408	25.5
Choi L., Liu Z., Matthews C.E., Buchowski M.S.	2011	Validation of accelerometer wear and nonwear time classification algorithm	Medicine and Science in Sports and Exercise	404	50.5
Thornton C., Hutter F., Hoos H.H., Leyton-Brown K.	2013	Auto-WEKA: Combined selection and hyperparameter optimisation of classification algorithms	Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	284	47.3
Kou G., Lu Y., Peng Y., Shi Y.	2012	Evaluation of classification algorithms using MCDM and rank correlation	International Journal of Information Technology and Decision Making	261	37.3

Otukei J.R., Blaschke T.	2010	Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms	International Journal of Applied Earth Observation and Geoinformation	261	29.0
Xia R., Zong C., Li S.	2011	Ensemble of feature sets and classification algorithms for sentiment classification	Information Sciences	260	32.5
Smits P.C., Dellepiane S.G., Schowengerdt R.A.	1999	Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach	International Journal of Remote Sensing	219	11.0
Stowe L.L., Davis P.A., Mcclain E.P.	1999	Scientific basis and initial evaluation of the CLAVR-1 global clear/cloud classification algorithm for the advanced very high-resolution radiometer	Journal of Atmospheric and Oceanic Technology	202	10.1
Kadah Y.M., Farag A.A., Zurada J.M., Badawi A.M., Youssef A.-B.M.	1996	Classification algorithms for quantitative tissue characterisation of diffuse liver disease from ultrasound images	IEEE Transactions on Medical Imaging	199	8.7
Park H.S., Ryzhkov A.V., Zrnić D.S., Kim K.-E.	2009	The hydrometeor classification algorithm for the polarimetric WSR-88D: Description and application to an MCS	Weather and Forecasting	193	19.3
Woo Thomas Y.C.	2000	Modular approach to packet classification: Algorithms and results	Proceedings - IEEE INFOCOM	190	10.0
Martin A.C.R., Thornton J.M.	1996	Structural families in loops of homologous proteins: Automatic classification, modelling and application to antibodies	Journal of Molecular Biology	190	8.3
Fei B., Liu J.	2006	Binary tree of SVM: A new fast multi-class training and classification algorithm	IEEE Transactions on Neural Networks	178	13.7
Brown I., Mues C.	2012	An experimental comparison of classification algorithms for imbalanced credit scoring data sets	Expert Systems with Applications	172	24.6
Lehmann C., Koenig T., Jelic V., Prichep L., John R.E., Wahlund L.-O., Dodge Y., Dierks T.	2007	Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG)	Journal of Neuroscience Methods	163	13.6

Figure 10 exhibits the network visualisation map analysis of the citations by documents with 20 minimum number of citations of a document in the classification field. [51] have a lot of connecting lines with multiple citations, demonstrating the document is being co-cited with multiple documents. Documents in the same cluster with similar colour are typically co-cited together.

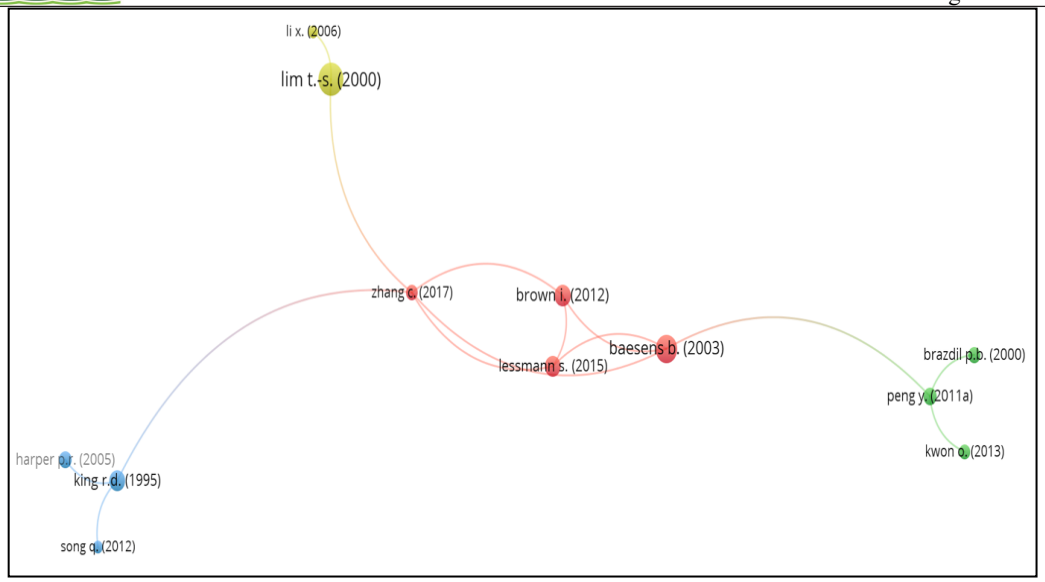

Fig. 10: Network visualisation map of the citations by documents with 20 minimum number of citations of a document

Meanwhile, the network visualisation map of the citations by active authors in classification-related publication with a minimum number of ten documents and ten citations were visualised in Figure 11. The map shows the authors who were engaged in classification research and received a higher number of citations. Closed circles suggested involved authors working closely together in the study. Some names probably may not be visible here due to the overlapping of names.

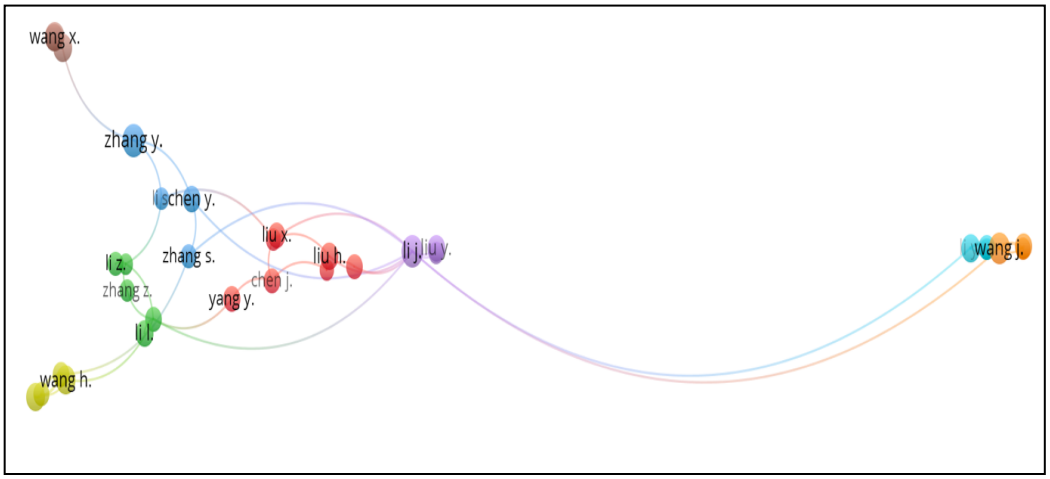

Fig. 11: Network visualisation map of the citations by authors with ten minimums

Network visualisation map of the citations by international collaboration among countries with five minimum number of documents and five minimum number of citations of a country (see Figure 12). The thickness of the connecting line between any two countries reveals the strength of citations by country. For instance, the strength of the link between China and the United States shows a thick line indicating they have strong citations collaboration. On the other hand, the line between China and Iraq shows a weak citation relationship as the connection line is thin. Countries that have similar colours demonstrate a single cluster. For example, countries with green colour like the Russian Federation, Canada, Egypt, Sweden and Serbia existed in a single cluster. The United Kingdom, Germany, Switzerland and Malaysia were clustered in red, and their link strength is a thick line with China. Hence, the bulk of their collaboration is with China.

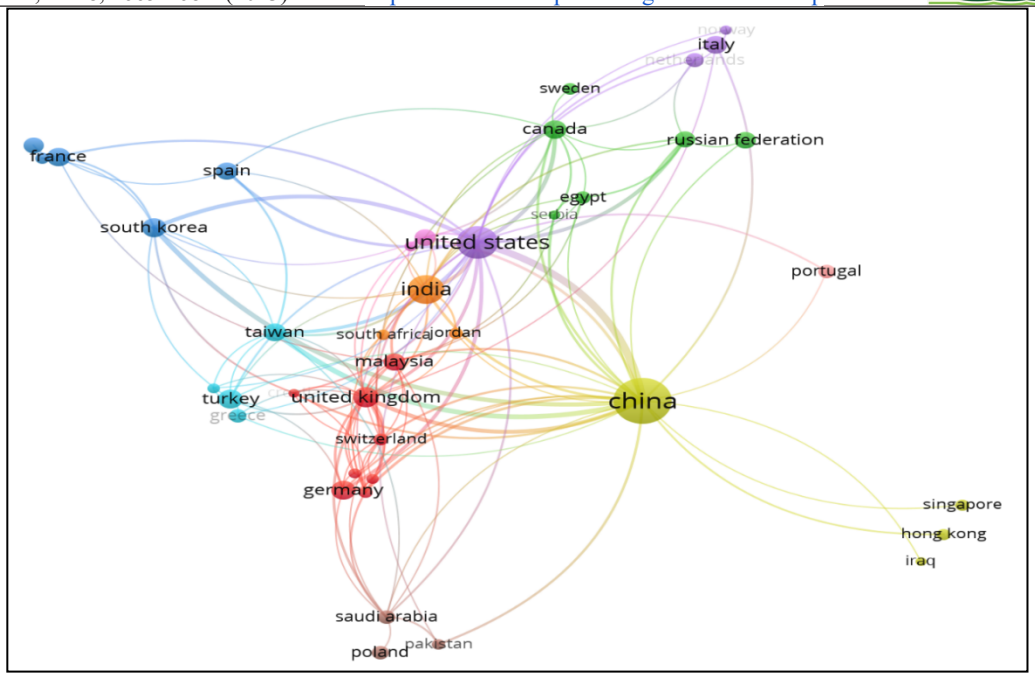

Fig. 12: Network visualisation map of the citations by country with five minimum number of documents and five minimum number of citations of a country

4 Scope and Limitations of Analysis

Whilst this research contributes research knowledge on the trend of literature related to classification modelling, there are a few restrictions that need to be understood. Firstly, the keywords used in the search query for this paper are only limited to literature that contains the specific terms in the title of the documents. Only four terms (i.e., either "classification modelling" OR "classification algorithm" OR "discriminant modelling" OR "discriminant algorithm") were used as search terms to write this article for the purpose of analysis. There are possibilities that some of the related literature used classification modelling in their studies but did not mention it in the title. They probably show it in the abstract, keywords, or just within the text in the documents. However, our focus of the study is clear when focusing only on the article title of the documents.

Secondly, the interpretation of a bibliometric map is not thoroughly straightforward. This is because bibliometric mapping has some restrictions, the interpretation of a map should always be made in a very cautious way. Basically, there are two restriction types of bibliometric mapping; boundaries imposed by the data and boundaries imposed by the map [52].

Thirdly, the bibliometric analysis is based solely on the Scopus database. Other databases are also available, and the outcomes may vary according to the database used (e.g., Google Scholar and Scopus) as well as the use of other search terms (e.g., classification rule or machine learning). The analysis included here is merely the articles with the availability of the author's keywords to convey the keywords network. The citation threshold with ">100" was employed and is denoted as the highly cited article in this paper. Due to these factors, all conclusions presented in this paper should be made within these restrictions' context.

5 Concluding Remarks

This paper has displayed the global research trends and scholarly networks on classification or discrimination research covering the period 1968-2020 retrieved from the Scopus database. The analysis comprised of bibliometric statistics regarding authors, author keywords, citations, journals, institutions and countries. Trends have been identified where it was conducted and published, sustainable and rising interest in classification, with the evident spread of thoughts globally and into specific topic areas. There were also significant changes in the types of articles published over time and its content. The applications of classification appear to be increasingly focused on Computer science-based research (i.e. neural network, decision tree, support vector machine, data mining, artificial intelligence) rather than discussing on its conceptual and theoretical contours, as well as lesser statistically based.

Conflicts of Interest Statement

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgment:

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UUM/02/5) with S/O code 14374.

References

[1] A. A. Olosunde and A. T. Soyinka. Discrimination and Classification of Poultry Feeds Data. *International Journal of Mathematical Research*, **2**(5), 37-41 (2013). DOI:10.1080/00207390600819 003

[2] D. J. Hand. Classifier Technology and the Illusion of Progress. *Statistical Science*, **21**(1), 1-15 (2006).

[3] J. N. Crook, D. B. Edelman and L. C. Thomas. Recent Developments in Consumer Credit Risk Assessment. European *Journal of Operational Research*, **183**, 1447-1465 (2007). DOI:10.1016/j.ejor.2006.09.100

[4] S. Banerjee and S. Pawar. Predicting Consumer Purchase Intention: A Discriminant Analysis Approach. *NMIMS Management Review*, **XXIII**, 113-129 (2013).

[5] M. L. Birzer and D. E. Craig-Moreland. Using Discriminant Analysis in Policing Research. Professional Issues in *Criminal Justice*, **3**(2), 33-48 (2008).

[6] Y. Guo, T. Hastie and R. Tibshirani. Regularised Linear Discriminant Analysis and Its Application in Microarrays. *Biostatistics*, **8**(1), 86-100 (2007).

[7] M. C. Carakostas, K. A. Gossett, G. E. Church and B. L. Cleghorn. Veterinary Pathology Online. *Veterinary Pathology*, **23**, 254-269 (1986).

[8] J. Y. Goulermas, A. H. Findlow, C. J. Nester, D. Howard and P. Bowker. Automated Design of Robust Discriminant Analysis Classifier for Foot Pressure Lesions using Kinematic Data. *IEEE Transactions on Biomedical Engineering*, **52**(9), 1549-1562 (2005).

[9] W. M. Maclaren. Using Discriminant Analysis to Predict Attacks of Complicated Pneumoconiosis in Coalworkers. *Journal of the Royal Statistical Society, Series D (The Statistician*), **34**(2), 197-208 (1985).

[10] Y. Takane, H. Bozdogan and T. Shibayama. *Ideal Point Discriminant Analysis. Psychometrika*, **52**(3), 371-392 (1987).

[11] M. J. Alrawashdeh, S. R. M. Sabri and M. T. Ismail. Robust Linear Discriminant Analysis with Financial Ratios in Special Intervals. *Applied Mathematical Sciences*, **6(**121), 6021-6034 (2012).

[12] E. I. Altman. Financial Ratios: Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, **23**(4), 589-609 (1968).

[13] R. A. Eisenbeis. Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics. *The Journal of Finance*, **32**(3), 875-900 (1977).

[14] L. J. Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao and S. H. Friend. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*, **415**(6871), 530-536 (2002).

[15] R. P. Hauser and D. Booth. Predicting Bankruptcy with Robust Logistic Regression. *Journal of Data Science*, **9**, 565-584 (2011).

[16] R. F. Engle and S. Manganelli. Quantile Prediction. In: G. Elliott & A. Timmermann (Eds.), *Economic Forecasting*, Elsevier, Netherland, pp: 964-968 (2004).

[17] T. Neideen and K. Brasel. Understanding Statistical Tests. *Journal of Surgical Education*, **64**, 93-96 (2007).

[18] D. Sheskin. *Handbook of Parametric and Non-parametric Statistical Procedures* (3rd ed.). Chapman and Hall/CRC,

United States, pp: 1193 (2004).

[19] M. Dumarey, B. Dejaegher, A. Durand and Y. V. Heyden. Exploratory Data Analysis and Classification of Capillary Electrophoretic Data. In: G. Hanrahan & F. A. Gomez (1st ed.), *Chemometric Methods in Capillary Electrophoresis*, pp: 291-321 (2009).

[20] W. J. Krzanowski. Discrimination and Classification using Both Binary and Continuous Variables. *Journal of American Statistical Association*, **70**(352), 782-790 (1975).

[21] W. J. Krzanowski. The Location Model for Mixtures of Categorical and Continuous Variables. *Journal of Classification*, **10**(1), 25-49 (1993).

[22] J. J. Higgins. *An Introduction to Modern Nonparametric Statistics*. Brooks/Cole, California, USA, pp: 366 (2004).

[23] Y. Yu. *Bayesian and Non-parametric Approaches to Missing Data Analysis*. Ph.D. dissertation, University of California, California, USA (2012).

[24] T. W. Kim. *Non-parametric Approaches for Drought Characterization and Forecasting*. B.A. thesis, University of Arizona, Arizona, USA (2003).

[25] M. H. Kim and P. A. Yoo. A Semiparametric Model Approach for Financial Bankruptcy Prediction. *IEEE International Conference on Engineering of Intelligent Systems*. New Jersey, pp: 1-6 (2006).

[26] X. Chen, O. Linton and V. I. Keilegom. Estimation of Semiparametric Models when the Criterion Function is Not Smooth. *Econometrica*, **71**, 1591-1608 (2003).

[27] D. J. Bauer. A Semiparametric Approach to Modeling Nonlinear Relations among Latent Variables. *Structural Equation Modeling*, **12**, 513-535 (2005).

[28] J. A. Anderson. Separate Sample Logistic Discrimination. *Biometrics*, **59**, 19-35 (1972).

[29] P. C. Chang and A. A. Afifi. Classification based on Dichotomous and Continuous Variables. *Journal of the American Statistical Association*, **69**(346), 336-339 (1974).

[30] I. G. Vlachonikolis and F. H. C. Marriott. Discrimination with Mixed Binary and Continuous Data. *Applied Statistics*, **31**(1), 23-31 (1982).

[31] W. J. Krzanowski. Selection of Variables, and Assessment of Their Performance, in Mixed-variable Discriminant Analysis. *Computational Statistics & Data Analysis*, **19**, 419-431 (1995).

[32] O. Asparoukhov and W. J. Krzanowski. Non-parametric Smoothing of the Location Model in Mixed Variable Discrimination. *Statistics and Computing*, **10**, 289-297 (2000).

[33] H. Hamid, L. M. Mei and S. S. S. Yahaya. New Discrimination Procedure of Location Model for Handling Large Categorical Variables. *Sains Malaysiana*, **46**(6), 1001-1010 (2017).

[34] H. Hamid, P. A. H. Ngu and F. M. Alipiah. New Smoothed Location Models Integrated with PCA and Two Types of MCA for Handling Large Number of Mixed Continuous and Binary Variables. *Pertanika Journal of Science & Technology*, **26**(1), 247-260 (2018).

[35] W. F. Massey. Principal Components Regression in Exploratory Statistical Research. *Journal of American Statistical Association*, **60**, 234-246 (1965).

[36] E. K. Kemsley. Discriminant Analysis of High-Dimensional Data: A Comparison of Principal Component Analysis and Partial Least Squares Data Reduction Methods. *Chemometrics and Intelligent Systems*, **33**, 47-61 (1996).

[37] H. Hamid. Winsorized and Smoothed Estimation of the Location Model in Mixed Variables Discrimination. *Applied Mathematics & Information Sciences: An International Journal*, **12**(1), 133-138 (2018a).

[38] H. Hamid. New Location Model based on Automatic Trimming and Smoothing Approaches. *Journal of Computational and Theoretical Nanoscience*, **15**, 493-499 (2018b).

[39] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**(2), 179-188 (1936).

[40] A. Ahmi and R. Mohamad. Bibliometric Analysis of Global Scientific Literature on Web Accessibility. *International Journal of Recent Technology and Engineering*, **7**(6), 250-258 (2019).

[41] W. M. Sweileh, S. W. Al-Jabi, A. S. AbuTaha, S. H. Zyoud, F. M. A. Anayah and A. F. Sawalha. Bibliometric Analysis of Worldwide Scientific Literature in Mobile - Health: 2006-2016. *BMC Medical Informatics and Decision*

Making, **17**(1), 72 (2017). https://doi.org/10.1186/s12911-017-0476-7

[42] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis and G. Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and Weaknesses. *The FASEB Journal*, **22**(2), 338-342 (2008).

[43] A. Ahmi and M. H. Mohd Nasir. Examining the Trend of the Research on Extensible Business Reporting Language (XBRL): A Bibliometric Review. *International Journal of Innovation, Creativity and Change*, **5**(2), 1145-1167 (2019).

[44] B. Fahimnia, J. Sarki and H. Davarzani. Green Supply Chain Management: A Review and Bibliometric Analysis. *Int J Prod Econ*., **62**, 101-114 (2015).

[45] W. M. Sweileh, S. W. Al-Jabi, A. S. AbuTaha, S. H. Zyoud, F. M. A. Anayah and A. F. Sawalha. Bibliometric Analysis of Worldwide Scientific Literature in Mobile - Health: 2006-2016. *BMC Medical Informatics and Decision Making*, **17**(1), 72 (2017). https://doi.org/10.1186/s12911-017-0476-7

[46] X. Zhang, R. C. Estoque, H. Xie, Y. Murayama and M. Ranagalage. Bibliometric Analysis of Highly Cited Articles on Ecosystem Services. *PLoS One*, **14**(2), p. e0210707 (2019).

[47] H. J. Li, H. Z. An, Y. Wang, J. C. Huang and X. Y. Gao. Evolutionary Features of Academic Articles Co-keyword Network and Keywords Co-occurrence Network: based on Two-mode Affiliation Network. *Phys. A*., **450**, 657-669 (2016).

[48] H. Liao, M. Tang, L. Luo, C. Li, F. Chiclana and X. J. Zeng. A Bibliometric Analysis and Visualisation of Medical Big Data Research. *Sustainability*, **10**(1), 166 (2018). https://doi.org/10.3390/su10010166

[49] D. X. Gu, J. J. Li, X. G. Li and C. Y. Liang. Visualising the Knowledge Structure and Evolution of Big Data Research in Healthcare Informatics. *Int. J. Med. Inform*., **98**, 22-32 (2017).

[50] E. Bauer and R. Kohavi. Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, **36**(1-2), 105-139 (1999).

[51] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, **54**(6), 627-635 (2003).

[52] R. Heersmink, J. van den Hoven, N. J. van Eck and J. van den Berg. Bibliometric Mapping of Computer and Information Ethics. *Ethics and information technology*, **13**(3), 241 (2011). DOI: 10.1007/s10676-011-9273-7