SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Visualising Current Research Trends in Class Imbalance using Clustering Approach: A Bibliometrics Analysis

Nurul Syahida Abu Bakar[1,2], Wan Fairos Wan Yaacob[3,4*], Yap Bee Wah[5], Wan Marhaini Wan Omar[6], Utriweni Mukhaiyar[7]

[1] School of Computer Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2] STEM Foundation Center, Universiti Malaysia Terengganu (UMT), 21300 Kuala Nerus, Terengganu, Malaysia
[3] Mathematical Sciences Studies, College of Computing, Informatics and Media, Universiti Teknologi MARA Cawangan Kelantan, Kampus Kota Bharu, 15050 Kota Bharu, Kelantan, Malaysia
[4] Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[5] UNITAR Graduate School, UNITAR International University, 47301 Petaling Jaya, Selangor, Malaysia
[6] Faculty of Business and Management, Universiti Teknologi MARA (UiTM), Cawangan Kelantan, Kampus Kota Bharu, 15050 Kota Bharu, Kelantan Malaysia
[7] Statistics Research Division, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Indonesia

ARTICLE INFO

Article history:
Received 15 June 2023
Received in revised form 10 December 2023
Accepted 26 December 2023
Available online 30 January 2024

Keywords:
Imbalance problem; clustering approach; bibliometric analysis; network analysis; science mapping; big data

ABSTRACT

In recent years, extensive research has been carried out on big data class imbalance problems using the clustering approach. The bibliometric analysis employs statistical techniques to map and assess trends in a specific research domain based on keywords, author affiliations, and citations. Bibliographic analysis assists us in comprehending unstructured big data. This study aims to present a comprehensive literature review on class imbalance problems using the clustering approach and identify gaps in the research domain using bibliometric analytical techniques. The Scopus and Web of Science databases were used to extract 663 articles on class imbalance data using a clustering approach published between 2010 and 2021. We used the VOS (Visualisation of Similarities) viewer to visualise the bibliometric analytical outcomes. Co-citation and co-word analysis were used to visualise the publication trend and identify areas of current research interest. The study's key findings evidenced a growing interest in the research domain. Herrera, f., and Chawla N. V. are dominant authors in this field, and China is leading the publication in the clustering approach for the big data imbalance problem. The top three affiliations are from China: Tsinghua University, the Chinese Academy of Sciences, and Beihang University. Conducting an in-depth bibliometric analysis using other databases such as Science Direct, IEEE, and Emerald is recommended. This study may assist researchers in understanding the nature of the big data imbalance problem using a clustering approach and providing insights for future research derived from these worldwide databases.

* *Corresponding author.*
E-mail address: wnfairos@uitm.edu.my

https://doi.org/10.37934/araset.38.2.95111

1. Introduction

Data mining and machine learning are becoming crucial for data-driven decision-making in the current era of big data technologies. Data mining algorithms can extract meaningful patterns from structured and unstructured data [1]. It entails the creation of data models from existing data sets to gain insights into a phenomenon pattern, relationship, and prediction of an event. The classification model is a robust machine-learning technique. The goal of classification is to predict instances based on independent attributes accurately. Overall performance accuracy is a criterion used to assess the classifier's effectiveness. The higher the accuracy, the better the classification is.

Dealing with imbalanced data sets is one of the challenges in the classification domain. Imbalanced data sets occur when one or more classes are underrepresented in the dataset. The imbalanced distribution is a common issue because the data instances of interest (such as fraud or cancer) are often rare in quantity. This problem is omnipresent in many real-world classification data sets, such as bone fracture prediction for osteoporosis prevention [2], customer churn prediction [3], and face re-identification [4]. A classifier could achieve a high accuracy rate in classifying the imbalanced dataset [5]. However, other performance indexes, such as Kappa statistics or sensitivity, can be very low. As a result, the accuracy rate is insufficient to assess a classifier's performance in imbalanced learning. This inaccuracy is due to inadequate information when using singular assessments like error rate or the overall accuracy rate in imbalanced learning. Furthermore, classifiers may treat the instances in the minority as outliers or simply ignore them. The classifier will have low sensitivity; thus, sampling and clustering methods have been proposed to address imbalance class problems.

The imbalance data issue affects the performance of classifiers where the model will be biased toward the majority class. Even though the classification accuracy may be high, the model will have zero or low sensitivity. This issue affects prediction accuracy and leads to bias in decision-making since this situation tends to skew rigorously toward the majority class. Various studies have been carried out to tackle the imbalance issue and provide a meaningful classification with zero misclassification error.

Japkowicz [6], identified the class imbalance problem, and it has received a great deal of attention from researchers and practitioners, resulting in an increasing number of publications on imbalanced learning. Due to the technological advancements in the twenty-first century, the amount of data generated doubles each year. The likelihood of encountering a significantly imbalanced dataset increases with data volume. A high-class imbalance, frequently observed in big data, complicates and challenges a learner's identification of the minority class because it introduces a bias in favour of the majority class. As a result, this becomes a tough task for the researcher to effectively distinguish between two classes (minority and majority), particularly in extremely imbalanced datasets.

Thus, this study conducted a bibliometric analysis using the clustering approach to discover trends in big data class imbalance in this research domain. Bibliometric analysis is a superior option compared with meta-analysis and systematic literature review since it can summarise large amounts of bibliographic data to present the current state of the intellectual structure and emerging trends in a research field. Additionally, the visualisation of similarities approach in this study is appropriate when the scope of the review is broad and the dataset is too large for manual review [7]. This study aims to identify the gap in the research domain by adopting bibliometric analytical techniques and to determine networks (i.e., the links between authors, countries, institutions, and journals) and research attention (i.e., change in stated research keywords). In this work, we described the methodology of visualising bibliometric analysis with analytical tools to map the current research trend in the big data problem of class imbalance using the clustering approach. The analysis included

the focal key terms, journal selection process, and visualisation of the similarity matrix. The Scopus analyser, WoS analyser, and VOSviewer tool were used for bibliometric analysis. We then present and discuss the results and the directions for future research.

The following is how this paper is structured: Section 2 provides an overview of bibliometric analysis literature focusing on big data imbalance problems using a clustering approach. The methodology used in the study is described in Section 3. Section 4 presents the results of the study and their in-depth discussion. Finally, Section 5 concludes the paper with some recommendations for future research.

2. Bibliometric Analysis

According to Hawkins [8], bibliometrics is "the quantitative investigation of the bibliometric characteristics of a body of literature." In other words, bibliometric analysis is a way of determining the influence of published papers and citations through statistical analysis. It aids researchers in determining the most active authors, institutions, nations, journals, and types of work within a topic, as well as the number of citations and author collaboration [9]. Bibliometrics analysis has gained prominence lately as it helps to identify trends in a research domain or topic [10-13].

Two aspects of bibliometric mapping that differ from the traditional structured review method are the development of bibliometric maps and their graphical display. In the bibliometric literature, creating bibliometric maps receives the most attention. Data on keyword co-occurrences can be used to generate so-called co-word maps, visual representations of a scientific field's structure [14]. The number of publications in which both keywords appear together in the title, abstract, or keyword list is the number of co-occurrences of two keywords [15]. In this technique, the article is considered the unit of analysis, whereas bibliometric mapping depicts interconnections among articles by showing how often an article is cited and co-cited by other articles [16].

Aside from co-word maps, bibliometric analysis using VOSviewer can generate co-citation analysis, which assumes that published articles in scholarly journals are based on their research on previously published articles [17]. In co-citation analysis, the degree to which researchers are cited in the same publication determines their relatedness. The greater the frequency with which two researchers are cited in the same publication, the closer they are related [18].

Many works in the machine learning literature have used bibliometric techniques [19], public health [20], and autoML in cash severity prediction [21]. However, there are limited studies on the clustering approach's bibliometric analysis of class imbalance problems. MARAŞ and Çiğdem [19], in 2022, was the first study of bibliometric analysis on imbalanced datasets. The research consisted of data collected from Scopus throughout 1957-2021 for articles that include keywords imbalanced dataset, imbalanced big dataset, oversampling, smote, unbalanced big dataset, unbalanced dataset, and under sampling. The keywords search focused on the sampling techniques, from which 16255 publications have been extracted, most of which cover the SMOTE and ensemble of SMOTE techniques. According to MARAŞ and Çiğdem [19], the imbalanced datasets problem is primarily studied by authors from the United States, China, and Germany. Nitesh V. Chawla and Francisco Herrera are the most cited first authors in the field. With recent development, imbalanced datasets in deep learning and big data have gained much attention. Although the emerging trend in the classification of imbalanced datasets with bibliometric analysis can be seen recently, there are still a few ongoing studies on the classification of imbalanced datasets using the clustering approach. Some bibliometric analyses cover topics on big data in agriculture [22,23], deep learning [24] and machine learning [25].

This study used a bibliometric technique to map class imbalance problem literary activity and assess the structure and patterns using quantitative methods in analysing journals, books, proceedings, and other publications. Scopus and WoS Analyzer were used to perform descriptive analysis on topics such as publication trends, top journals in terms of publications, affiliation and country statistics, and popular authors. Co-authorship network, co-citation, and co-word analysis were visualised using VOSViewer software.

3. Methodology

3.1 Database Choice and Searching Strategy

The literature search was conducted in August 2022 to form a database of sources. Scopus and Web of Science (WoS) database were selected as it is the world's largest and most comprehensive abstracts and citation database, covering a wide domain of publications in science, technology, medicine, social sciences, and the arts and humanities. The following keywords of "Class Imbalance and Clustering" (("class imbalance") AND ("clustering")) OR "imbalance dataset and clustering" (("imbalance dataset") AND ("clustering")) were used in the search for the titles, abstracts, and keyword fields of the database. Only publications from 2010 to 2021 were considered for this study since the number of articles published before 2010 is between 1-3, with no significant trend that could be assessed. The preliminary search yielded 288 articles from the Scopus database search and 375 articles from the WoS database search. These articles focused on class imbalance problems using a clustering approach and were restricted to the English language, irrespective of subject areas and source types. All 663 articles were reviewed during the screening stage based on the title and abstract relevancy. Articles were chosen using a macro approach (top-down), which began with a broad search based on article titles and progressed to an investigation of the relevance abstract using a clustering approach in an imbalanced dataset. As a result, 77 Scopus articles and 147 WoS articles were retained for further analysis. Each source of information extracted data elements such as the abstract, author name, year of publication, author keywords, author affiliation, name of the institution, name of the journal, and the number of citations.

The article selection process is summarised in Figure 1. The data was exported to an Excel spreadsheet. Scopus Analyzer and VOSviewer were used to analyse the data. Scopus and WoS Analyzer were used to perform the descriptive analysis. The publication patterns, the most prolific journals, contributing authors, institutions, nations, and the most cited articles were all tracked using an Excel spreadsheet, Scopus, and WoS Analyzer for 12 years (2010 to 2021). The retrieved CSV text files were exported to VOSviewer to construct a bibliographic map and bibliometric network. VOSviewer was also used for co-citation and co-word analysis to explore research patterns and clusters in the field of study [15].

3.2 Analytical Tool

The main analysis tool used to map research trends in class imbalance problems with the clustering approach is VOSviewer version 1.6.18, based on bibliometrics. Van Eck and Waltman developed VOSviewer in the Java-based computer program that can construct, visualise, and explore node-link maps based on bibliographic data [26,27]. VOS is an abbreviation for Visualising Similarities, a mapping technique for computer programmers. Particularly, VOSviewer generates bibliometric maps through the three steps outlined below [27]:

i. A similarity matrix is obtained by normalizing the co-occurrence matrix. In which VOSwiever uses association strength as a similarity measure. Thus, the similarity s_{ij} between two items in the co-occurrence data is calculated using association strength as

$$s_{ij} = \frac{c_{ij}}{w_i w_j} \tag{1}$$

c_{ij} denotes the number of co-occurrences between items *i* and *j*, while w_i and w_j denote the total number of occurrences or co-occurrences of items *i* and *j*, respectively.

ii. The VOS technique is applied to the previous step's similarity matrix. This step aims to minimise the weighted sum of all pairs of elements squared Euclidean distances

$$\min V(x_1,\ldots,x_n) = \sum_{i<j} s_{ij} \left\| x_i - x_j \right\|^2, \quad \text{subject to } \frac{2}{n(n-1)} \sum_{i<j} \left\| x_i - x_j \right\| = 1 \tag{2}$$

where $x_i = (x_{i1}, x_{i2})$ denotes the location of item *i* in a two-dimensional map, ||•|| denotes the Euclidean norm, and *n* denotes the number of items to be mapped.

iii. Transform the solutions of Eq. (2) by translation, rotation, and reflection to ensure that VOSviewer produces consistent results.

These three steps of the similarity matrix index are performed directly in the VOSwiever based on the predetermined threshold.

The node-link map created by VOSviewer represents a bibliometric network of an object. Authors, articles, countries, affiliations, and keywords are the objects in this context. Each link has association strength represented by a positive numerical value. The greater the value of association strength, the stronger the connection between the objects. A high value indicates a strong connection. As a result, these maps are used to analyse a discipline's research trends visually. The overall summary of the methodology for Bibliometric Analytical Techniques is shown in Figure 1.

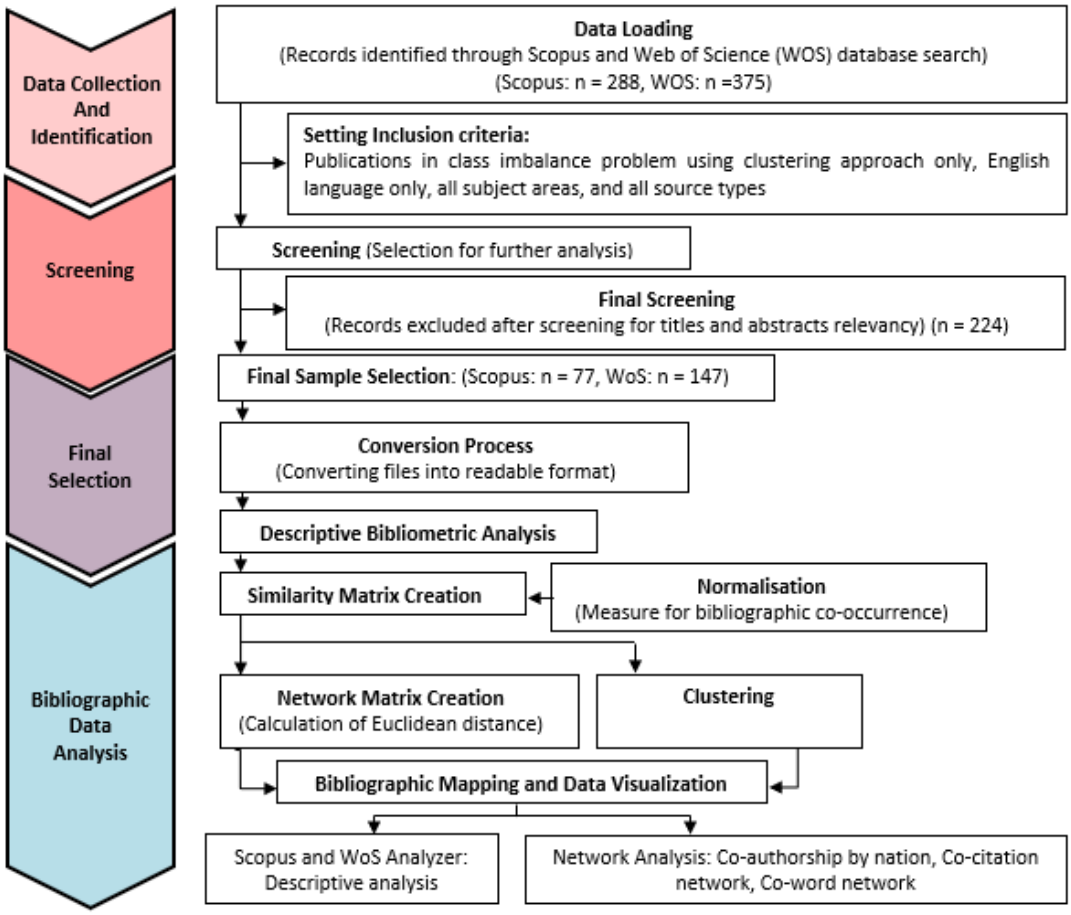

Fig. 1. Methodology for bibliometric analytical techniques

4. Result and Discussion

4.1 Description Analysis

The number of academic articles on the subject is the most direct indicator of publication trends, and this number can adequately reflect the research process in a specific field. This section describes the distribution of article publications on imbalanced datasets using cluster analysis over time. Figure 2 depicts the steadily rising publication trend from 2010 to 2021. Results showed that the publication was stagnant between 2010 to 2016, with only about 3 to 10 articles published. However, the number of related articles increased from 5 in 2000 to 54 in 2021. In the last 4 years (2017-2021), 180 articles have been published, constituting 80.3% of the total publications (224 articles) in this field. The hike in publications indicates that the application of clustering in addressing imbalanced datasets began to expand rapidly after 2016. The research outputs increased sharply from 2016 to 2017. There were 224 articles published, and the trend over the years can be divided into two phases: the initial phase from 2010 to 2016 (which accounts for 19.6% of all articles published); and the growth phase from 2017 to 2021.

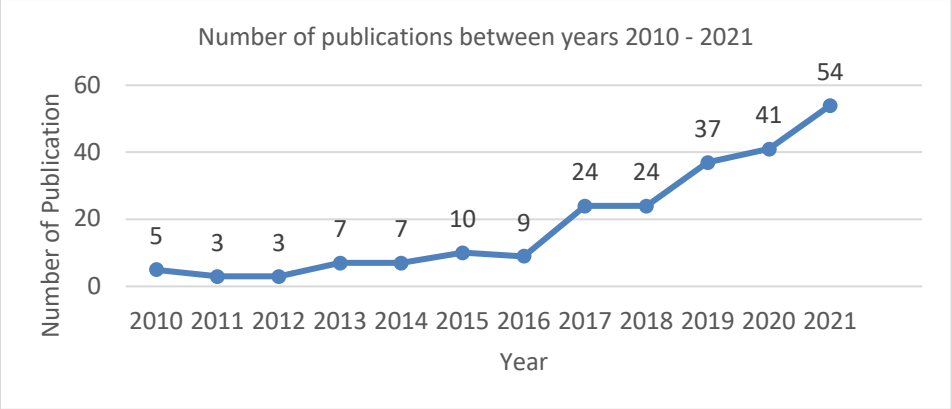

Fig. 2. Publication trend of imbalance problem with clustering approach

Figure 3 shows that when the number of publications in Scopus and WoS databases is compared, the WoS database contributes more published articles than Scopus, except in 2012. In general, WoS lead the publications in this field than Scopus. The growing number of publications reflects the academic interest in the clustering method for dealing with imbalanced datasets. Furthermore, the trend also suggests that there will be a further increase in the number of publications.

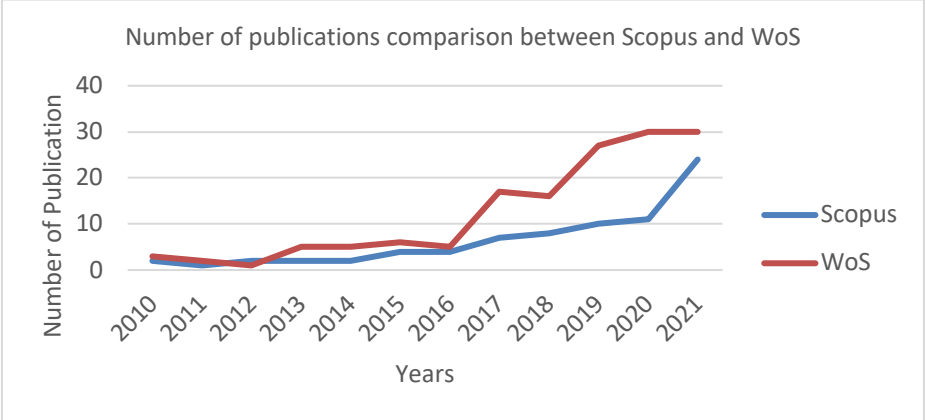

Fig. 3. Comparison of publication trend between Scopus and WoS

4.2 Journals

Next, we determine the top journals on this research topic. The 224 retrieved articles were published in 138 different journals. Among these 138 journals, 113 (58%) published only one article, 26 (12%) published two, 15 (7%) published three, and the remaining 53 (24%) published more than three articles. Table 1 lists the top four journals with the most publications in this field, and the journals are sorted by the number of articles published. The Information Sciences received the most citations (715), although it only published 8 articles in this field. On the contrary, Lecture Notes in Computer Science, which included subseries lecture notes in artificial intelligence bioinformatics, ranked first for the number of articles published (10), but only 25 citations. The journal that receives the highest number of citations is ranked as the top journal. Therefore, Information Sciences, IEEE Access, and Expert System with Applications were identified as the top-performing journals.

Table 1
Top three journals in imbalance problem with clustering approach publication

Journal	Impact factor	Number of citations	Number of articles	Percentage (%)
Lecture Notes in Computer Science, including subseries lecture notes in artificial intelligence and bioinformatics	2.1	25	10	5%
IEEE Access	3.367	146	9	4%
Information Sciences	5.524	715	8	4%
Expert Systems with Applications	8.665	102	7	3%

4.3 Countries and Affiliations

Table 2 depicts the geographic distribution of the top five countries contributing to the imbalance problem with clustering approach publications. China contributes the most to publications, with 90 (36%) articles, followed by India, with 34 (14%) articles. These two countries predominate the publication in this field. Nonetheless, the distribution of countries' publications worldwide still indicates that this field of study receives global attention.

According to MARAŞ and Çiğdem [19], the imbalanced dataset problem is primarily studied by authors from the United States, China, and Germany. However, in this study, China and India appear to be the most productive countries for publishing imbalanced datasets using the clustering approach.

Table 2
Top 5 countries contributing to the topic of the imbalance problem with a clustering approach

Location	China	India	USA	Taiwan	South Korea
Number of articles	90	34	20	13	12
Percentage (%)	36%	14%	8%	5%	5%

To further analyse the country's co-authorship in this field, a visualised network map created by VOSviewer for the collaboration between the countries is displayed in Figure 4. The node size on the map represents a country's influence in this field, while the thickness of links indicates cooperative closeness among different countries. Based on Table 2, the three most productive countries are China, India, and USA. Among these three countries, collaboration only exists between China and USA. Despite being second in the ranking, India has no collaboration with other countries.

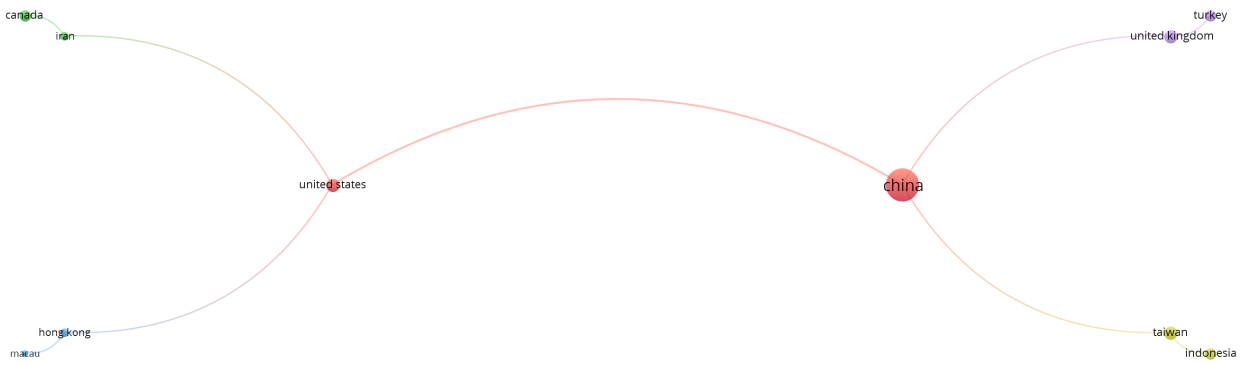

Fig. 4. Countries who work together

We also look at the top research affiliations in this field. Based on Scopus and WoS Analyzer, the list of the top four institutions that published the most papers on imbalance problems with the clustering approach, geographic location, number of articles published, and number of citations is shown in Table 3. State University System of Florida and Beihang University rank first and second, respectively, in terms of the number of articles published. However, these two affiliations have fewer citations (20 and 25, respectively) than Tsinghua University, which has 45 citations despite only publishing four articles. Therefore, it is noticeable that Tsinghua University is more influential. Of the 4 top affiliations, 3 come from China, implying that China significantly impacts this field.

Table 3
Top four affiliations contributing to the imbalance problem with a clustering approach

Affiliations	Country	Number of articles	Number of citations
State University System of Florida	US	5	20
Beihang University	China	5	25
Chinese Academy of Sciences	China	4	35
Tsinghua University	China	4	45

4.4 Co-Authors Analyses

In this section of co-author's citation analysis, the unit of analysis is co-authors. The authors in this study were chosen based on a minimum of ten citations. Table 4 and Figure 5 show the prominent co-citation authors and network visualisation in this field, respectively. A total of 99 authors out of 3274 authors met the threshold and were chosen for co-citation network analysis. The total link strength of authors was calculated for each of the 99 authors. However, only the top six total link strengths were reported in Table 4, and those with fewer than 50 citations were excluded from this study. Herrera, F. was identified as the author with the greatest total link strength of 4131 with the highest citations of 102. Chawla, N.V. came second with 3307 total link strength and 96 citations. The other three authors, Fernandez, A., Hall, L.O., and Japkowicz, N., have fewer than 100 citations with total link strength of 2900, 1989, and 1971, respectively. In this study, the prominent co-citations are Herrera, f. and Chawla N. V., supported by [18], since these two authors emerge as the most cited first authors.

The size of the nodes represents the normalised number of citations received by articles, while the thickness of the line represents the strength of co-citation ties. Moreover, the colours of the nodes indicate the identified cluster to which the article belongs.

Table 4
The five prominent authors with the highest total link strength

Authors	Total link strength	Citations
Herrera, F.	4131	102
Chawla, N. V.	3307	96
Fernandez, A.	2900	72
Hall, L. O.	1989	57
Japkowicz, N.	1971	56

As shown in Figure 5, VOSViewer developed 4 clusters where the cluster marked in red was the strongest cluster with 37 items, followed by a green colour cluster which consisted of 32 items. A blue cluster with 29 items was identified as the third strongest cluster, and the weakest cluster was a yellow cluster with only 1 item. As a result, the authors actively cited in this field can be found in the clusters mentioned above.

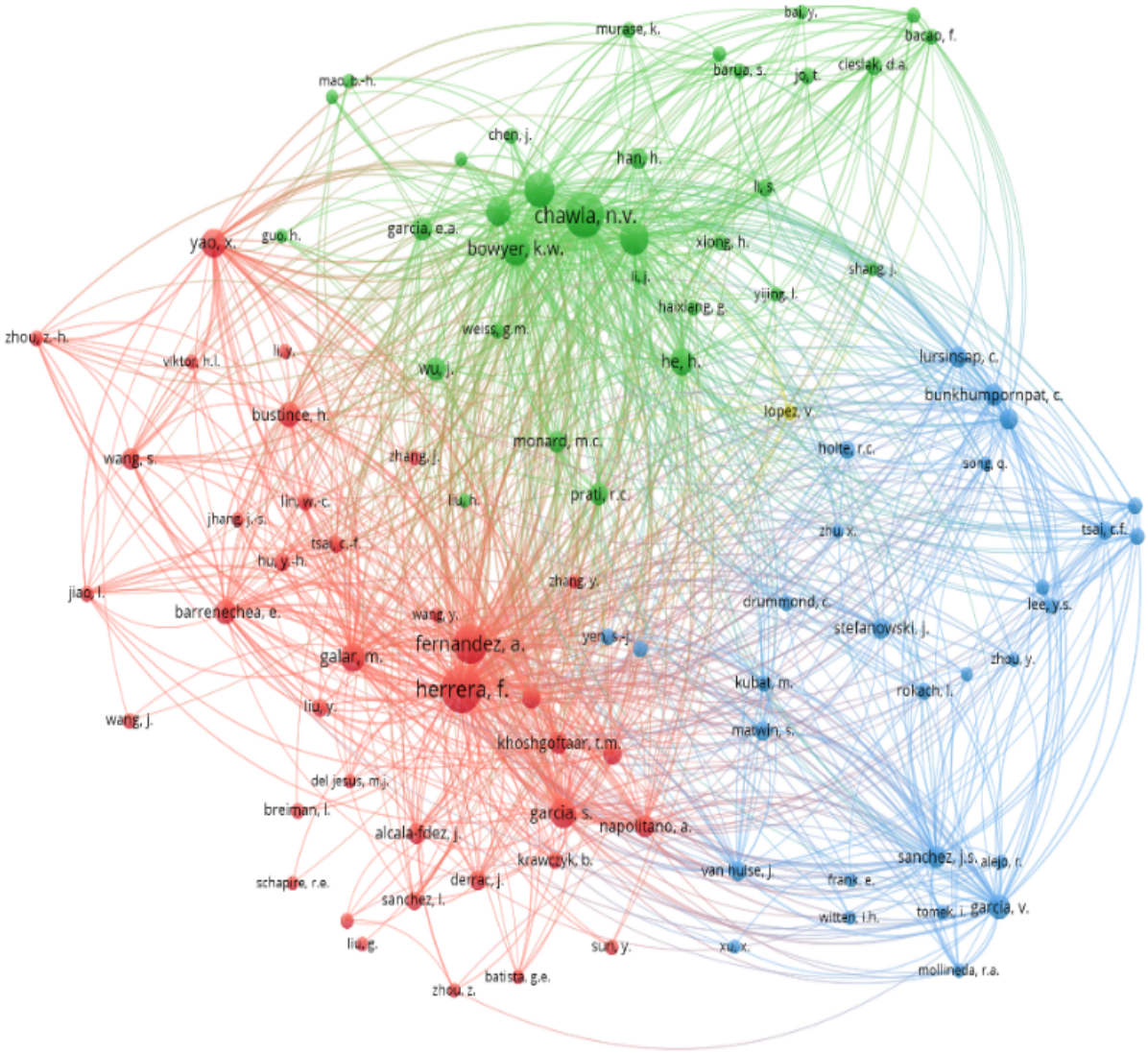

Fig. 5. Visualising the Co-citation Network

4.5 Co-Citation Analysis

Citations are used as a measurement of academic influence. If a publication or an author has high citations, the publication or author is considered influential in the field. The most highly cited articles on the imbalance problem with the cluster approach are shown in Table 5. The year of publication, total citations, the title, the authors, and technique/impact are listed for each article.

Table 5
Most cited articles on imbalance problems with cluster approach

Year	Total citation	Title	Authors	Technique / Impact
2017	335	Clustering-based undersampling in class-imbalanced data	[28]	Clustering-based undersampling approach with the nearest neighbours of the cluster centres outperformed five state-of-the-art approaches.
2018	290	Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE	[29]	The oversampling method based on k-means and SMOTE outperforms other popular oversampling methods.
2019	117	Under-sampling class imbalanced datasets by combining clustering analysis and instance selection	[30]	The Cluster-Based Instance Selection (CBIS) method is introduced, which combines clustering analysis and instance selection. CBIS perform significantly better than the other six state-of-the-art with Bagging and boosting-based MLP ensemble as classifiers.
2021	100	An ensemble machine learning model based on multiple filtering and supervised attribute clustering algorithm for classifying cancer samples	[31]	Multiple Filtering and Supervised Attribute Clustering algorithms were proposed based on the Ensemble Classification model (MFSAC-EC). The proposed method is significantly better than high-dimensional microarray gene expression datasets.
2021	92	Robust Vehicle Classification Based on Deep Features Learning	[32]	Semi-Supervised Fuzzy C-Mean (SSFCM) clustering was discussed. SSFCM can reduce the sensitivity of unsupervised fuzzy C-means (FCM) clustering algorithm and improve the classification performance when dealing with the multi-class imbalanced dataset.
2017	81	Evolutionary Cluster-Based Synthetic Oversampling Ensemble (ECO-Ensemble) for Imbalance Learning	[33]	Cluster-based oversampling ensemble framework outperforms current state-of-the-art ensemble algorithms by tackling class imbalance problems.
2017	77	Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem	[34]	Fast-CBUS and the Pareto frontier are superior in both computational cost and performance.
2014	74	Cluster-based sampling of multi-class imbalanced data	[35]	Clustering with sampling for Multi-class Imbalanced classification using Ensemble (C-MIEN) achieves higher performance than state-of-the-art methods.
2019	70	Semi-Supervised Deep Fuzzy C-Mean Clustering for Imbalanced Multi-Class Classification	[36]	Semi-supervised deep Fuzzy C-mean clustering for imbalanced multi-class classification (DFCM-MC) could perform better due to their ability to recognise and consolidate fundamental information from unsupervised data.
2020	65	A Boosting-Aided Adaptive Cluster-Based Undersampling Approach for Treatment of Class Imbalance Problem	[37]	Boosting aided adaptive cluster-based undersampling was discussed.

The most influential article was written by Lin W.C, Tsai C. F., Hu Y. H., and Jhang J. S. in 2017 with 355 citations and published in the information science journal. The second and third-highest citations were 269 and 107 times, respectively. These two articles received a high number of citations as their publications were relatively recent in 2018 and 2019. All the authors of the third-highest article are also authors of the first article, except for Yao G. -T.

The top 10 articles, as presented in Table 5, employed clustering methods involving nearest neighbours, k-means, affinity propagation, and fuzzy C-means. Four of the ten papers used method

clustering with undersampling [28,30,34,37], and two used method clustering with oversampling [29,33]. Moreover, two of the top three studies conducted clustering with undersampling [28,30] and one applied clustering with oversampling. When implementing the clustering strategy, clustering with undersampling is the most frequently employed approach for addressing imbalance issues.

4.6 Co-Word Analysis

Co-word analysis was applied to show the relationship among the keywords in each field [38]. A keyword is the unit of co-word analysis. The threshold should be established based on the number of times the keywords appear in a document [39]. A typical mapping of co-word analysis begins by importing a text file from the Scopus database into VOSviewer. VOSViewer has a powerful user graphic interface to generate maps easily [40,41]. Meanwhile, VOS mapping is used in VOSviewer to generate a two-dimensional diagram that depicts the location of two elements based on their distance [40].

Based on the 224 articles, all keywords were extracted to generate a map of the co-word network, as shown in Figure 6. The threshold was set to a minimum recurrence of 9 times, resulting in 28 keywords meeting this criterion. Figure 6 depicts the relationships between the keywords in this research area. It is important to note that the thickness of the line demonstrates the strength of the keyword co-occurrence. The elements on the map's edges have a small number of links, whereas the central location indicates strong relationships with a larger number of other keyword groups [42]. VOSviewer generated five clusters using the cluster approach, as shown in Figure 6. The keywords were divided into clusters based on their frequent co-occurrence in specific Scopus-indexed articles. Figure 6 depicts five distinct colours: red, green, blue, yellow, and purple, corresponding to clusters 1, 2, 3, 4, and 5, respectively. Each bubble represents a keyword, and the size of each bubble is proportional to the co-occurrence frequencies of keywords.

Figure 6 shows that the strongest keyword is imbalanced data with a big red bubble, linked to more diverse groups of other keywords or frequently appeared in imbalanced problem studies.

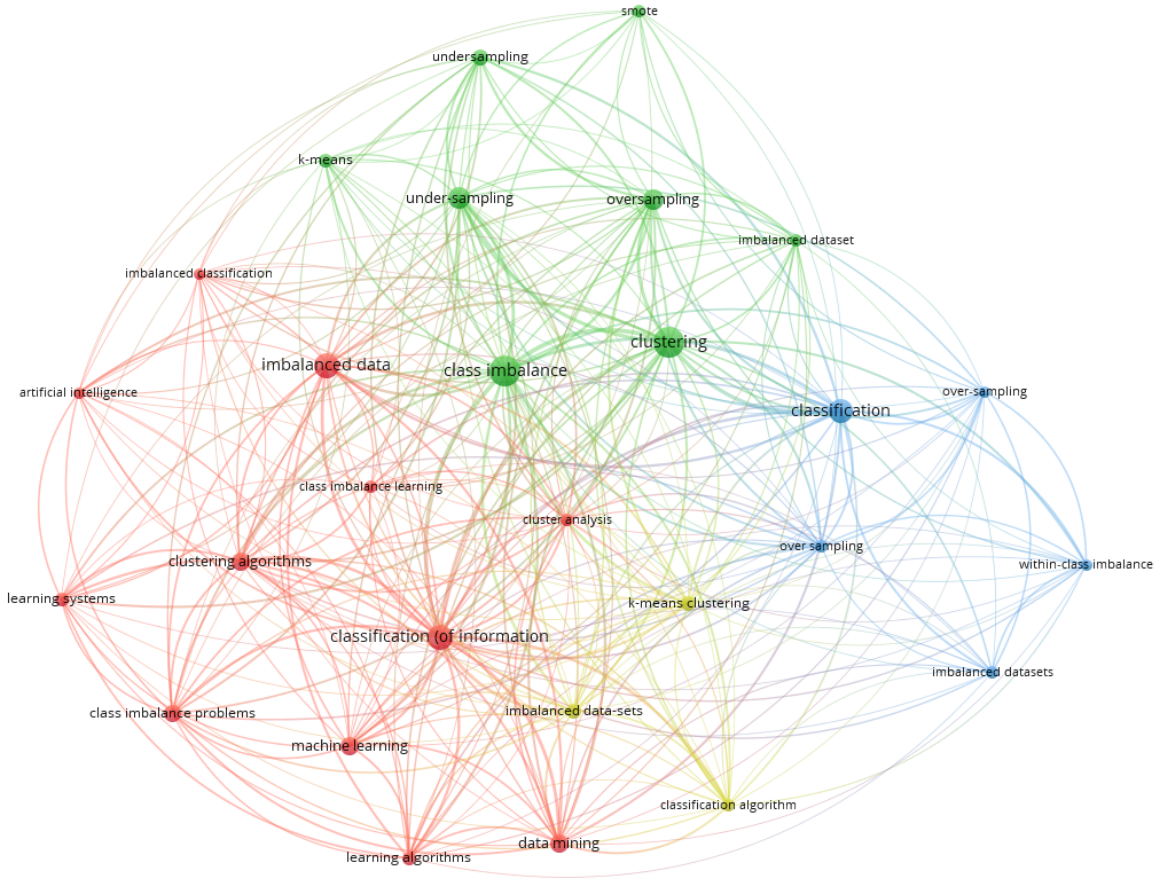

Fig. 6. Visualising co-word network on imbalance problem with a clustering approach

This fact is further corroborated by Table 6, which shows that it appeared 41 times. The second-most common keyword is clustering, which appeared 64 times in Table 6 and is depicted in Figure 6 with a green colour centre.

Table 6
Most frequently used keywords on imbalance problems with the clustering approach

Keywords	Frequency
class imbalance	63
clustering	62
Classification (of information)	42
Imbalanced data	41
classification	38
Under-sampling	32
oversampling	29
Clustering algorithm	24
Data mining	23
Machine learning	23

Table 7 lists all the keywords for the five clusters. Cluster 1 with 12 items, 'classification (of information)', 'imbalanced data', 'clustering algorithms', 'data mining', 'machine learning', 'class imbalance problems', 'learning systems', 'learning algorithm', 'cluster analysis', 'class imbalance learning', 'artificial intelligence', and 'Imbalance classification', are the most frequently used

keywords. Meanwhile, Cluster 3 consists of 'classification', 'imbalanced datasets', 'over sampling', 'over-sampling', and 'within-class imbalance'. Cluster 1 and 3 are merged to create a group that primarily concentrates on classification, encompassing class imbalanced data and data mining techniques. The second group comes from Cluster 2 and 4, which consists of articles mainly about techniques to deal with imbalance issues such as clustering, undersampling, oversampling and k-means clustering, with the following keywords appearing frequently: 'class imbalance', 'clustering', 'under-sampling', 'oversampling', 'undersampling', 'k-means', 'imbalance dataset', 'SMOTE', 'imbalanced datasets', 'k-means clustering', and 'classification algorithm'.

Table 7
Keywords for the four clusters on imbalance problems with a clustering approach

Cluster 1 (12 Items)	Frequency	Total link strength	Cluster 2 (8 Items)	Frequency	Total link strength
classification (of information)	42	204	class imbalance	63	209
imbalanced data	41	123	clustering	62	192
clustering algorithms	24	103	under-sampling	32	109
data mining	23	112	oversampling	29	104
machine learning	23	74	undersampling	17	70
class imbalance problems	19	84	k-means	13	33
learning systems	14	61	imbalance dataset	11	48
learning algorithm	13	77	SMOTE	11	28
cluster analysis	12	66			
class imbalance learning	11	37			
artificial intelligence	9	42			
Imbalance classification	9	34			
Cluster 3 (5 Items)	**Frequency**	**Total link strength**	**Cluster 4 (3 Items)**	**Frequency**	**Total link strength**
classification	38	138	imbalanced datasets	16	86
imbalanced datasets	11	36	k-means clustering	16	73
over sampling	11	79	classification algorithm	12	67
over-sampling	10	49			
within-class imbalance	9	42			

Table 8 summarises them into two groups of research focus.

Table 8
Research focus of four clusters in the imbalance problem with clustering approach

Groups	Number of items	Research focuses
Cluster 1 and 3	17	Class imbalance data and data mining techniques, which could be referred to as classification
Cluster 2 and 4	11	Techniques to deal with imbalance issues include clustering, undersampling, oversampling, and k-means clustering

5. Conclusion

This study applied bibliometric analysis to visualise scientific research on big data class imbalance problems with a clustering approach. A total of 224 out of 663 articles were extracted and analysed. The publication trends in big data class imbalance with clustering approach are increasing. China has the highest number of publications. Information Science, IEEE Access, and Expert Systems with Applications are the leading journals in this field. The top four leading institutions in addressing big data class imbalance problems are Tsinghua University, the Chinese Academy of Sciences, Beihang

University, and the State University System of Florida. Since the top three affiliations are from the same country, it is consistent with China being the top country in this field.

According to this study, China and India dominate the publications of imbalanced datasets using the clustering approach, yet there is no cooperation between them. In contrast, China and the United States have collaborative research. In this study, the prominent co-citations are Herrera, f. and Chawla N. V., and both of them are neither from China nor India.

The clustering methods used in the top 10 articles were nearest neighbours, k-means, affinity propagation, and fuzzy C-means. Four of the ten papers used the clustering method with undersampling [28,30,34,37], and two used the clustering method with oversampling [29,33]. The top three articles [28-30] conducted clustering with undersampling and with oversampling. Clustering with undersampling is the most often used method for dealing with imbalance problems when performing the clustering strategy.

Bibliometric analysis is useful for mapping research trends, article citations, and keyword analysis. Future research may consider performing bibliometric analysis on the imbalance problem using a clustering approach of other databases such as Science Direct, Emerald, and Elsevier. In terms of searching for titles, abstracts, and keywords in the database, future researchers might use the top keywords obtained in this study, such as 'imbalanced data' and 'clustering algorithms,' rather than 'Class Imbalance and Clustering' and 'imbalance dataset and clustering' that used in this study. Incorporating this aspect could increase the number of articles, thereby enhancing the accumulation of additional insights. This study provides insight into the general research trends in big data class imbalance using a clustering approach of bibliographic analysis.

Acknowledgement
This research was not funded by any grant.

References

[1] Mazurowski, Maciej A., Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, and Georgia D. Tourassi. "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance." *Neural networks* 21, no. 2-3 (2008): 427-436. https://doi.org/10.1016/j.neunet.2007.12.031

[2] Bach, Malgorzata, Aleksandra Werner, J. Żywiec, and Wojciech Pluskiewicz. "The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis." *Information Sciences* 384 (2017): 174-190. https://doi.org/10.1016/j.ins.2016.09.038

[3] Zhu, Bing, Bart Baesens, and Seppe KLM vanden Broucke. "An empirical comparison of techniques for the class imbalance problem in churn prediction." *Information sciences* 408 (2017): 84-99. https://doi.org/10.1016/j.ins.2017.04.015

[4] Soleymani, Roghayeh, Eric Granger, and Giorgio Fumera. "Progressive boosting for class imbalance and its application to face re-identification." *Expert Systems with Applications* 101 (2018): 271-291. https://doi.org/10.1016/j.eswa.2018.01.023

[5] Li, Jinyan, Simon Fong, Sabah Mohammed, and Jinan Fiaidhi. "Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms." *The Journal of Supercomputing* 72, no. 10 (2016): 3708-3728. https://doi.org/10.1007/s11227-015-1541-6

[6] Japkowicz, Nathalie. "Learning from imbalanced data sets: a comparison of various strategies." In *AAAI workshop on learning from imbalanced data sets*, vol. 68, pp. 10-15. AAAI Press Menlo Park, 2000.

[7] Donthu, Naveen, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. "How to conduct a bibliometric analysis: An overview and guidelines." *Journal of business research* 133 (2021): 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

[8] Hawkins, Donald T. "Bibliometrics of electronic journals in information science." *Information Research* 7, no. 1 (2001): 7-1.

[9] Blažun, Helena, Peter Kokol, and Janez Vošner. "Research literature production on nursing competences from 1981 till 2012: A bibliometric snapshot." *Nurse Education Today* 35, no. 5 (2015): 673-679. https://doi.org/10.1016/j.nedt.2015.01.002

[10] Pitt, Christine, Andrew Park, and Ian P. McCarthy. "A bibliographic analysis of 20 years of research on innovation and new product development in technology and innovation management (TIM) journals." *Journal of Engineering and Technology Management* 61 (2021): 101632. https://doi.org/10.1016/j.jengtecman.2021.101632

[11] Fan, Jingchun, Ya Gao, Na Zhao, Runjing Dai, Hailiang Zhang, Xiaoyan Feng, Guoxiu Shi *et al.,* "Bibliometric analysis on COVID-19: a comparison of research between English and Chinese studies." *Frontiers in public health* 8 (2020): 477. https://doi.org/10.3389/fpubh.2020.00477

[12] Shamsuddin, Jamaltul Nizam, Christopher Gan, and Dao Le Trang Anh. "Bibliometric Analysis of InsurTech." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 30, no. 2 (2023): 103-132. https://doi.org/10.37934/araset.30.2.103132

[13] Tan, Huiyi, Keng Yinn Wong, Hong Yee Kek, Kee Quen Lee, Haslinda Mohamed Kamar, Wai Shin Ho, Hooi Siang Kang *et al.,* "Small-scale botanical in enhancing indoor air quality: A bibliometric analysis (2011-2020) and short review." *Progress in Energy and Environment* (2022): 13-37. https://doi.org/10.37934/progee.19.1.1337

[14] Eck, Nees Jan van, and Ludo Waltman. "How to normalize cooccurrence data? An analysis of some well-known similarity measures." *Journal of the American society for information science and technology* 60, no. 8 (2009): 1635-1651. https://doi.org/10.1002/asi.21075

[15] Van Eck, Nees Jan, and Ludo Waltman. "Visualizing bibliometric networks." In *Measuring scholarly impact: Methods and practice*, pp. 285-320. Cham: Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-10377-8_13

[16] Van Eck, Nees Jan, and Ludo Waltman. "VOSviewer manual." *Leiden: Univeristeit Leiden* 1, no. 1 (2013): 1-53.

[17] van Raan, Anthony FJ. "Properties of journal impact in relation to bibliometric research group performance indicators." *Scientometrics* 92, no. 2 (2012): 457-469. https://doi.org/10.1007/s11192-012-0747-0

[18] Perianes-Rodriguez, Antonio, Ludo Waltman, and Nees Jan Van Eck. "Constructing bibliometric networks: A comparison between full and fractional counting." *Journal of informetrics* 10, no. 4 (2016): 1178-1195. https://doi.org/10.1016/j.joi.2016.10.006

[19] MARAŞ, Abdullah, and E. R. O. L. Çiğdem. "Emerging Trends in Classification with Imbalanced Datasets: A Bibliometric Analysis of Progression." *Bilişim Teknolojileri Dergisi* 15, no. 3 (2022): 275-288. https://doi.org/10.17671/gazibtd.1019015

[20] dos Santos, Bruno Samways, Maria Teresinha Arns Steiner, Amanda Trojan Fenerich, and Rafael Henrique Palma Lima. "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018." *Computers & Industrial Engineering* 138 (2019): 106120. https://doi.org/10.1016/j.cie.2019.106120

[21] Angarita-Zapata, Juan S., Gina Maestre-Gongora, and Jenny Fajardo Calderín. "A bibliometric analysis and benchmark of machine learning and automl in crash severity prediction: The case study of three colombian cities." *sensors* 21, no. 24 (2021): 8401. https://doi.org/10.3390/s21248401

[22] Kamilaris, Andreas, Andreas Kartakoullis, and Francesc X. Prenafeta-Boldú. "A review on the practice of big data analysis in agriculture." *Computers and Electronics in Agriculture* 143 (2017): 23-37. https://doi.org/10.1016/j.compag.2017.09.037

[23] Couliably, Solemane, Bernard Kamsu-Foguem, Dantouma Kamissoko, and Daouda Traore. "Deep learning for precision agriculture: A bibliometric analysis." *Intelligent Systems with Applications* (2022): 200102. https://doi.org/10.1016/j.iswa.2022.200102

[24] Li, Yang, Zeshui Xu, Xinxin Wang, and Xizhao Wang. "A bibliometric analysis on deep learning during 2007–2019." *International Journal of Machine Learning and Cybernetics* 11 (2020): 2807-2826. https://doi.org/10.1007/s13042-020-01152-0

[25] Su, Miao, Hui Peng, and Shaofan Li. "A visualized bibliometric analysis of mapping research trends of machine learning in engineering (MLE)." *Expert Systems with Applications* 186 (2021): 115728. https://doi.org/10.1016/j.eswa.2021.115728

[26] Van Eck, Nees Jan, and Ludo Waltman. "Citation-based clustering of publications using CitNetExplorer and VOSviewer." *Scientometrics* 111 (2017): 1053-1070. https://doi.org/10.1007/s11192-017-2300-7

[27] Van Eck, Nees, and Ludo Waltman. "Software survey: VOSviewer, a computer program for bibliometric mapping." *scientometrics* 84, no. 2 (2010): 523-538. https://doi.org/10.1007/s11192-009-0146-3

[28] Lin, Wei-Chao, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. "Clustering-based undersampling in class-imbalanced data." *Information Sciences* 409 (2017): 17-26. https://doi.org/10.1016/j.ins.2017.05.008

[29] Douzas, Georgios, Fernando Bacao, and Felix Last. "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE." *Information Sciences* 465 (2018): 1-20. https://doi.org/10.1016/j.ins.2018.06.056
[30] Tsai, Chih-Fong, Wei-Chao Lin, Ya-Han Hu, and Guan-Ting Yao. "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection." *Information Sciences* 477 (2019): 47-54. https://doi.org/10.1016/j.ins.2018.10.029
[31] Bose, Shilpi, Chandra Das, Abhik Banerjee, Kuntal Ghosh, Matangini Chattopadhyay, Samiran Chattopadhyay, and Aishwarya Barik. "An ensemble machine learning model based on multiple filtering and supervised attribute clustering algorithm for classifying cancer samples." *PeerJ Computer Science* 7 (2021): e671. https://doi.org/10.7717/peerj-cs.671
[32] Niroomand, Naghmeh, Christian Bach, and Miriam Elser. "Robust vehicle classification based on deep features learning." *IEEE Access* 9 (2021): 95675-95685. https://doi.org/10.1109/ACCESS.2021.3094366
[33] Lim, Pin, Chi Keong Goh, and Kay Chen Tan. "Evolutionary cluster-based synthetic oversampling ensemble (eco-ensemble) for imbalance learning." *IEEE transactions on cybernetics* 47, no. 9 (2016): 2850-2861. https://doi.org/10.1109/TCYB.2016.2579658
[34] Ofek, Nir, Lior Rokach, Roni Stern, and Asaf Shabtai. "Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem." *Neurocomputing* 243 (2017): 88-102. https://doi.org/10.1016/j.neucom.2017.03.011
[35] Prachuabsupakij, Wanthanee, and Nuanwan Soonthornphisaj. "Cluster-based sampling of multiclass imbalanced data." *Intelligent Data Analysis* 18, no. 6 (2014): 1109-1135. https://doi.org/10.3233/IDA-140687
[36] Arshad, Ali, Saman Riaz, and Licheng Jiao. "Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification." *IEEE Access* 7 (2019): 28100-28112. https://doi.org/10.1109/ACCESS.2019.2901860
[37] Devi, Debashree, Suyel Namasudra, and Seifedine Kadry. "A boosting-aided adaptive cluster-based undersampling approach for treatment of class imbalance problem." *International Journal of Data Warehousing and Mining (IJDWM)* 16, no. 3 (2020): 60-86. https://doi.org/10.4018/IJDWM.2020070104
[38] Leung, Xi Y., Jie Sun, and Billy Bai. "Bibliometrics of social media research: A co-citation and co-word analysis." *International Journal of Hospitality Management* 66 (2017): 35-45. https://doi.org/10.1016/j.ijhm.2017.06.012
[39] Zupic, Ivan, and Tomaž Čater. "Bibliometric methods in management and organization." *Organizational research methods* 18, no. 3 (2015): 429-472. https://doi.org/10.1177/1094428114562629
[40] Cobo, Manuel J., Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. "Science mapping software tools: Review, analysis, and cooperative study among tools." *Journal of the American Society for information Science and Technology* 62, no. 7 (2011): 1382-1402. https://doi.org/10.1002/asi.21525
[41] Feng, Yunting, Qinghua Zhu, and Kee-Hung Lai. "Corporate social responsibility for supply chain management: A literature review and bibliometric analysis." *Journal of Cleaner Production* 158 (2017): 296-307. https://doi.org/10.1016/j.jclepro.2017.05.018
[42] Lulewicz-Sas, Agata. "Corporate social responsibility in the light of management science–bibliometric analysis." *Procedia Engineering* 182 (2017): 412-417. https://doi.org/10.1016/j.proeng.2017.03.124