Journal of Scientometric Research, 2024; 13(1):43-57.
https://www.jscires.org

Research Article

Bibliometric Analysis of Recent Trends in Machine Learning for Online Credit Card Fraud Detection

Dickson Hove*, Oludayo Olugbara, Alveen Singh

MICT SETA 4IR Center of Excellence, Department of Information Technology, Durban University of Technology, Berea, Durban, SOUTH AFRICA.

ABSTRACT

Online credit card fraud (OCCF) is the malicious act of using credit card details belonging to another person to complete fraudulent transactions over the Internet. Naturally, masses of researchers have engaged in the imperative search for effective solutions across a wide range of disciplines. The result is a rich tapestry of methodologies, models, frameworks, and inventions exhibiting dramatic spread and growth. However, this also results in an unorganized research domain. In this state, a bibliometric analysis is a useful technique for establishing a reconciled snapshot of the OCCF research domain. This paper has particular interest in determining the intellectual structure of the knowledge of machine learning, deep learning, and ensemble learning models for early detection of OCCF. This bibliometric analysis is conducted using 524 publications between 2013 and 2022 extracted from the SCOPUS core collection database. Microsoft Excel, VOSViewer, and Biblioshiny software tools were used for data analysis. The findings indicate that ensemble learning models are trending and the three most authoritative authors have been exposed in this study. There is a sharp rise in global publications annually and India has the most publications with the most impactful authors. Five broad clusters of knowledge are imbalanced data, anomaly detection, machine learning, decision trees, and ensemble learning. Intellectual collaboration across regions is strong amongst Asia, Europe, and North America with weak associations between Africa and South America. This is the first bibliometric analysis in the domain of OCCF detection to the best of the author's ability. The findings significantly contribute to the application of OCCF detection through the creation of intellectual patterns in existing literature. The results bring about synthesis within a domain of research that is currently disorganized. This in turn helps researchers to identify research gaps, and areas for further research and formulate a curriculum.

Keywords: Bibliometric analysis, Credit card, Deep learning, Ensemble learning, Machine learning, Online fraud.

Correspondence:
Dickson Hove
MICT SETA 4IR Center of Excellence,
Department of Information Technology,
Durban University of Technology,
Berea, Durban, SOUTH AFRICA.
Email: hovedickson@gmail.com
ORCID: 0000-0003-4005-8939

Received: 08-12-2022;
Revised: 24-08-2023;
Accepted: 03-02-2024.

INTRODUCTION

Credit card fraud in the simplest definition is the unauthorised use of a credit card by an unknown entity to conduct a financial transaction. Online Credit Card Fraud (OCCF) occurs exclusively on the internet, a malicious act of unlawfully purchasing goods and services online using another person's credit card details.[1-3] This misdemeanour results in financial loss and sometimes psychological trauma for the card owner, poor publicity for the financial service provider, and encourages ever bolder attempts at online fraud. It is a contemporary and urgent global socio-economic challenge as a new digital economy proliferates at an unbridled rate. Financial service providers devise and adopt numerous technological and computational-driven prevention and detection schemes however, the evolving sophistication of tools and strategies to conduct fraud continues to evade these schemes and result in increasingly massive financial losses.

Anti-fraud schemes in OCCF can be categorized into prevention, detection, and containment.[4] Prevention includes pre-emptive steps to reduce the occurrence of OCCF.[5,6] These encompass activities devoted to sensitization and behavioural change among card owners as well as computer-assisted security tools for analysis of previous OCCF. Detection requires more sophisticated computation tools that can accurately separate a fraudulent activity from an acceptable one in real-time or in near real-time periods.[3,7] Containment strategies include blocking further usage of credit cards once they have been compromised.[8] From the three broad categories, detection is of concern in this paper and more specifically, the detection of OCCF.

Extant literature presents a broad spectrum of techniques for detection. For some time, developments in AI have been viewed by researchers as promising avenues for the increasingly complex

DOI: 10.5530/jscires.13.1.4

OCCF detection problem. At the time of writing, a Google Scholar search using simple keywords online credit card fraud detection+machine learning returns 17,000 related publications since 2018. Traditional rule-based Machine Learning (ML), Deep Learning (DL) and Ensemble Learning (EL) feature extensively in the returned results and are of particular interest to this paper.

ML solves problems by building a model that is an accurate representation of a certain multidimensional dataset.[9] This is achieved by equipping the ML model with the computational ability to find statistically derived patterns and relationships within the given dataset.[10] In the context of OCCF detection, a ML model aids in identifying usage patterns in legacy data which denotes fraudulent activity. In ML, two further categorisations can be made namely, supervised ML or unsupervised ML.[11]

In supervised ML, OCCF detection aims to decide whether or not a transaction is fraudulent based on legacy and labelled data.[12] Popular supervised ML algorithms in use include data mining techniques, Logistic Regression (LR), and Support Vector Machines (SVM) working with certain rules in support of a human expert on a labelled dataset.[13,14] However, unsupervised ML based on the OCCF detection system identifies suspicious usage patterns in transaction logs where illegitimate and genuine transactions coexist.[15] The dataset is not necessarily labelled when implementing the unsupervised ML algorithms. OCCF detection is achieved by using an anomaly or outlier approach to extract unusual user behavior profiles based on their historical transaction records and use it as a base to verify if an incoming transaction is fraudulent.[16,17]

As a mature research area, DL has received much interest and experimentation in the area of OCCF detection. DL is a form of ML that enables computational systems to learn from experience and understand the world in terms of a hierarchy of concepts.[18-21] DL is the next evolutionary step from the family of ML techniques that allows computational models to learn by example on their own. Instead of grooming a given dataset to run through predefined algorithms, DL can set up parameters about the dataset and train the computational system to learn on its own by recognizing patterns using many layers of processing. There are various variants of DL models, which include but are not limited to, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), adversarial Neural Networks (NN), and deep autoencoders [22-25] each of which are frequently applied to the OCCF detection research domain.

EL is being extensively experimented with in OCCF detection research.[26] EL is a combination of two or more ML algorithms to solve a computational problem better than would have been done when the algorithms were trained in silos.[27-30] Due to the technological advancements, the domain is now well associated with big data rendering traditional rule-based less effective when dealing with huge datasets.[31] Several algorithms such as autoencoders, Artificial Neural Networks (ANN), SVM, and LR have been proposed to deal with different aspects of OCCF detection, such as feature engineering, imbalanced data, and concept drift.[32,33] These challenges can be effectively resolved by combining more than one algorithm to create an EL approach.[31] Bagging, stacking, and boosting[34] present as the more popular categories in EL. The flexibility and adaptability of two or more ML techniques have led to the emergence of EL to achieve improved performance in a computational model. EL has been proposed in the OCCF detection space[28,34-36] with commendable results in the literature.

ML, DL, and EL drive the more recent inventions in the OCCF detection problem. Within the confines of this triad, there is a wide dispersion and growth of the topic. Despite our best efforts, there are a few recent attempts at rigorous synthesis of extant literature. Those publications that did attempt synthesis of extant literature derive results from manual literature review, content, or concept analysis, which are not unanimously well known for scientific rigour. These results make a firm grasp of directional impetus and research gaps a difficult endeavour. The current condition presents several subsequent problems for advancing knowledge in the area of OCCF, establishing needed agendas for future research, and developing relevant teaching materials.

A bibliometric analysis in this domain is desirable to determine these trends and the evolution of knowledge from ML, DL, and EL in OCCF detection. Bibliometric analysis is a statistical approach used to quantitatively investigate a set of scientific outputs and track their evolution over time.[37] Bibliometric analysis has been effectively employed in the examination of healthcare research, wastewater treatment studies, and renewable energy studies.[38-40] As of 06 November 2022, an advanced search on the SCOPUS database using keywords (TITLE-ABS-KEY ("online credit card fraud") AND TITLE-ABS-KEY ("bibliometric analysis")) produced zero results. Hence, as of this date, and to the best of our search efforts, no bibliometric analysis has been conducted to summarise the OCCF detection research domain. Therefore, this paper sets out to provide a rigorous overview of the literature on ML, DL, and EL applications in OCCF detection. To achieve this, a bibliometric analysis is performed guided by the following research questions:

RQ1: What is the annual trajectory in the number of publications?

RQ2: Which authors have been the most productive over the past ten years?

RQ3: What are the key knowledge concepts and clusters?

RQ4: What are the distribution of publications and the nature of collaboration across global geographic regions?

The analysis provides a snapshot of the OCCF detection research domain in the past 10 years. In turn, this analysis could benefit current and future researchers by unveiling the intellectual

landscape of knowledge. The next section describes the methods used to conduct this study. Section 1 introduces the study, section 2 describes the research methods. This is followed by section 3 on the findings and discussion and lastly, section 4 presents the conclusion.

METHODOLOGY

OCCF detection is a research domain characterized by an overwhelming number of publications and resultant bibliographic data. In this condition, a bibliometric analysis is desirable for the comprehension and analysis of such a vast body of knowledge. Bibliometric analysis endeavors toward rigorous analysis and visualization of a knowledge domain. Various parameters such as journal name, author, keywords, and country of publication can be interpreted from analysis of publications in the domain. Generally, a bibliometric analysis involves seven stages which are, data retrieval, pre-processing, network extraction, normalization, mapping, analysis, and visualization.[41] However, the steps are mostly implemented in parallel, for example, network extraction, normalization, mapping, analysis, and visualization are executed at the same time on given parameters. Therefore, the seven steps could be collapsed into only three steps namely, data identification, data extraction, and data analysis.[42] The following subsections describe the adopted bibliometric analysis procedure.

Searching and identification of data

Web of Science (WoS), PubMed, Dimensions, and SCOPUS are among the more popular online databases used to gather bibliometric analysis data for research purposes. Compared to the other databases, SCOPUS provides a larger citation and abstract search facility for peer-reviewed publications, most of the publications in SCOPUS are available in WoS but several publications in Scopus are not available in WoS, for example. Further, OCCF scholarly falls into the physical sciences space hence, SCOPUS was chosen specifically over other databases because it contains more than 96% of the peer-reviewed publications in physical sciences. The resources utilised in this study encompass scholarly articles spanning the period from 2013 to 2022, during which the research on this subject matter commenced to garner attention and interest. The study papers were collected using a comprehensive sample strategy to conduct a subsequent bibliometric analysis. The dataset includes bibliometric elements that are analysed, such as the publication title, author, abstract, keywords, publication year, publisher journal, type of publication, and affiliations. The study utilised search strings: "machine learning," OR "ensemble learning," OR "deep learning," AND "credit card fraud," to extract relevant material from the SCOPUS database. Consideration was paid to the contextual relevance of these search terms within the study. The inquiry underwent multiple iterations to incorporate the terms "online credit card fraud detection" and "online credit card fraud prevention". The inclusion criteria for this study were confined to articles, reviews, and conference proceeding papers that were in their final phases of publication. Additionally, only materials written in the English language were considered. Subsequently, 672 documents were identified.

This search strategy favours publications on ML, DL, and EL in OCCF detection. Conference proceedings, journal articles, and article reviews written in English were considered. Therefore 524 documents relating to OCCF detection were eligible for the final bibliometric analysis. Preferred Reporting Items For Meta-Analysis (PRISMA)[43] were utilized in selecting and screening documents for final bibliometric analysis as shown in Figure 1.

Analysis of data

The OpenRefine tool was used to eliminate duplicates, merge similar records, and fill in missing information within the included publications. OpenRefine is a free open-source data transformation and visualization tool.[44] The eligible documents from OpenRefine are further analysed using two popular tools namely, Bibliometrix and VOSViewer. These two software tools were used to form an integration of first and second-order

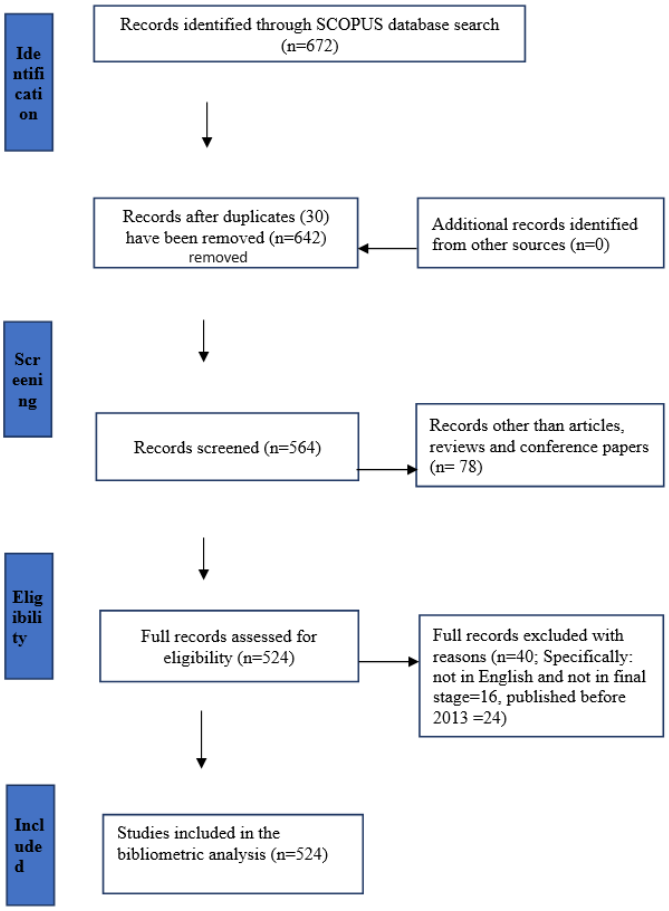

Figure 1: Flow chart diagram of the study selection method using PRISMA guidelines.

Note: Adapted from guidelines[43] PRISMA=Preferred reporting items for systematic reviews and meta-analyses; *n*=number of articles.

indicators to produce an objective bibliometric analysis and avoid a high probability of bias. In addition, Microsoft Excel 2019 was utilized to calculate the percentages and frequencies of each publication and to generate the corresponding graphs. Using Bibliometrix and VOSViewer, 524 documents that were published in 333 sources, from 2013 to 2022 each with an average citation of 12,08 were incorporated in this analysis. The analysis involves 1551 authors and an international co-authorship of 14.31%. The publications are stratified as 164 articles, 347 conference papers, and 13 reviews.

FINDINGS AND DISCUSSION

The aim of this study is a bibliometric analysis of literature on OCCF detection from 2013 to 2022 using 524 publications retrieved from the SCOPUS database. The research questions were formulated to objectify the bibliometric analysis which includes discovering the trajectory of the research field in several publications, the most authoritative authors and publishers, emerging topics, and collaboration strengths among regions. This section presents the findings of this analysis according to the study questions outlined in the introduction section.

Annual trajectory in published documents volume

Figure 2 depicts the number of publications on OCCF from 2013 to 2022 in response to the first research question. The number of articles produced from year to year is considered a reliable indicator of the research trend in the field of OCCF detection. Careful analysis of patterns in the number of publications can provide insight into the likely future direction of research. Figure 2 depicts a plot of the number of publications and cumulative publications on a year-on-year basis. However, researchers have demonstrated a great deal of interest in the domain since 2017. Post-2017, at least 10 documents have been published, as opposed to fewer than 8 each year before 2017. The greatest number of publications occurred in 2021, a total of 144. In 2022, by November 6, 2022, 115 publications had been achieved. In addition, the cumulative publications graph shows a sharp upward trajectory since 2017.

Table 1 displays ten primary sources in the categories of most relevant and most locally cited sources in which the 524 documents for this study have been published. It offers a thorough examination of the journal sources and the documents contained within them, emphasising their noteworthy contribution to the comprehension of OCCF detection. The colour green was used to signify the distribution and impact of the five most relevant and locally referenced sources among the top ten sources where the research materials were published. The publications collectively demonstrate their substantial contribution to the field of OCCF and may serve as valuable primary sources for both current and future researchers on this subject.

In addition, citation scattering network analysis was performed on the 524 papers to identify the weight of each document. The technique entails computing the number of citation links for each document and displaying the documents with the highest number of links as depicted in Figure 3. Nevertheless, a portion of the 524 documents in the network are not interconnected, whereas the largest group of interconnected documents comprises 270 documents. Consequently, the documents with larger dots exhibit a greater significance in terms of linkages and citations.

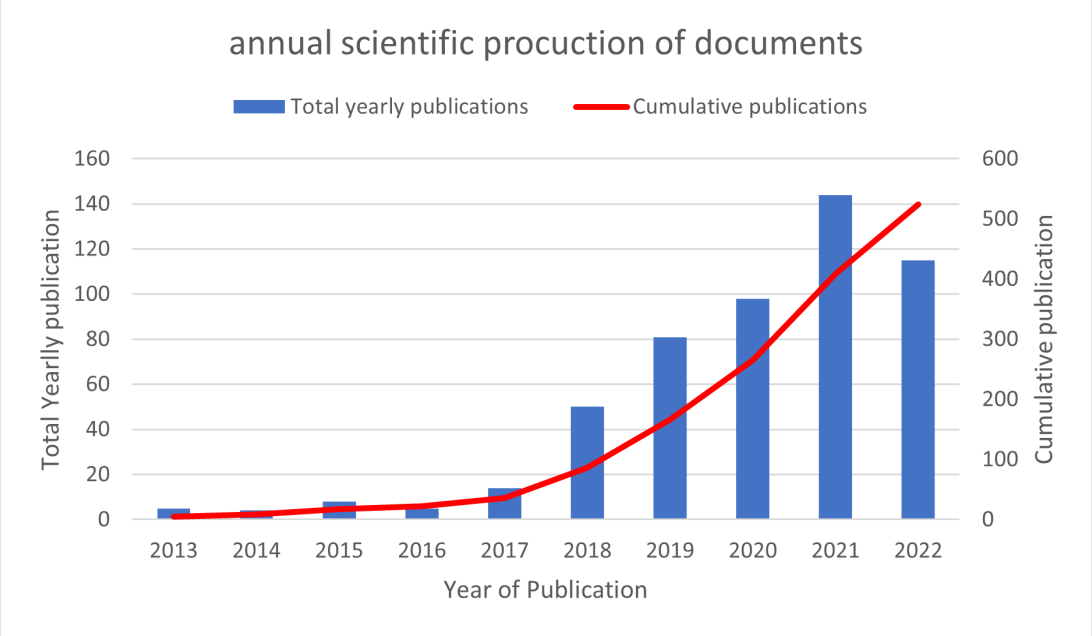

Figure 2: Annual scientific production of publications.

Most productive authors over the past 10 years

This study examines the co-citation networks of authors in the OCCF detection domain, with a focus on the cited references, sources, and publications. Figure 4 illustrates the reference co-citation network among authors, as visualised using the VOSViewer tool. Authors who did not have at least 10 citations were excluded from the analysis. For the authors who reached this threshold, the total co-citation strength was determined, as illustrated in Figure 4. Figure 4 depicts nodes that represent references, with the size of the node indicating the cumulative number of citations that have referenced the given citation. A co-citation relationship is symbolised by a link connecting two nodes, with the thickness of the link indicating the number of citations associated with the reference.

Subsequently, it was deemed necessary to analyse the links between authors and co-authors to identify the prominent research groups operating within a specific field. The accomplishment was attained by the process of delineating the connections between the authors and co-authors, as illustrated in Figure 5. The mapping process provides a visual representation of the relationship, which makes it easier to study not only the activity of an author but, the interconnectivity with other research groups as well. Figure 5 depicts that there are four major groups, considering at least three authors in a group. The research group of authors Zhang. Y, Chen. H, Cheng. D, Wang. X, Liu. C, Alzahrani. A, Baabdullah. T, li. Y, Hu. X and Wu. M was a prominent group in the research field. Among all the researchers in this group, Wang. X and Zhang. Y has the highest number of publications, 5 each, the rest of the 9 authors in this group have 4 publications each. The research group of Zhang. Y and Liu. G are highly connected and form the largest cluster of 19 researchers. Zhang. Y has a total link strength of 7 which means that the author was involved in collaborative research with 7 authors namely Hu. X, Wu. M, Cheng. D, Wang. X, Liu. G, Jiang. C and Yan. C.

However, in all the groups Caelene. O is amongst the most productive authors with 9 publications and a total link strength of 32. The author has collaborated with Lucas. Y, Calabretto. S, Granitzer. M, Portier. P, He-guelton. I, Bontempi. G, Carcillo. F, and Oble. F in online CCF detection research. Caelene. O falls in the second largest group consisting of 14 researchers. He-guelton. I and Caelene. O as mentioned before produced the highest number of publications in this group with 9 apiece, followed by Bontempi. G with 8 publications. Based on the information in Figure 4, Lui. G with 11 publications He-guelton. I and Caelen. O, 9 publications apiece are the most productive authors in this group.

All authors shown in Figure 4 are clustered into one of 5 color nodes, red, yellow, green, blue, and purple. Each color node represents a separate theme. The authors in the same color node have a strong relationship because they exhibit the same theme in their publications. In general, these 5 distinct schools of thought in the domain may be summarized as follows, red is imbalanced data, green is anomaly detection, purple is machine learning, blue is centered on decision trees, and yellow focuses on ensemble learning. Further, authors are depicted through a network of

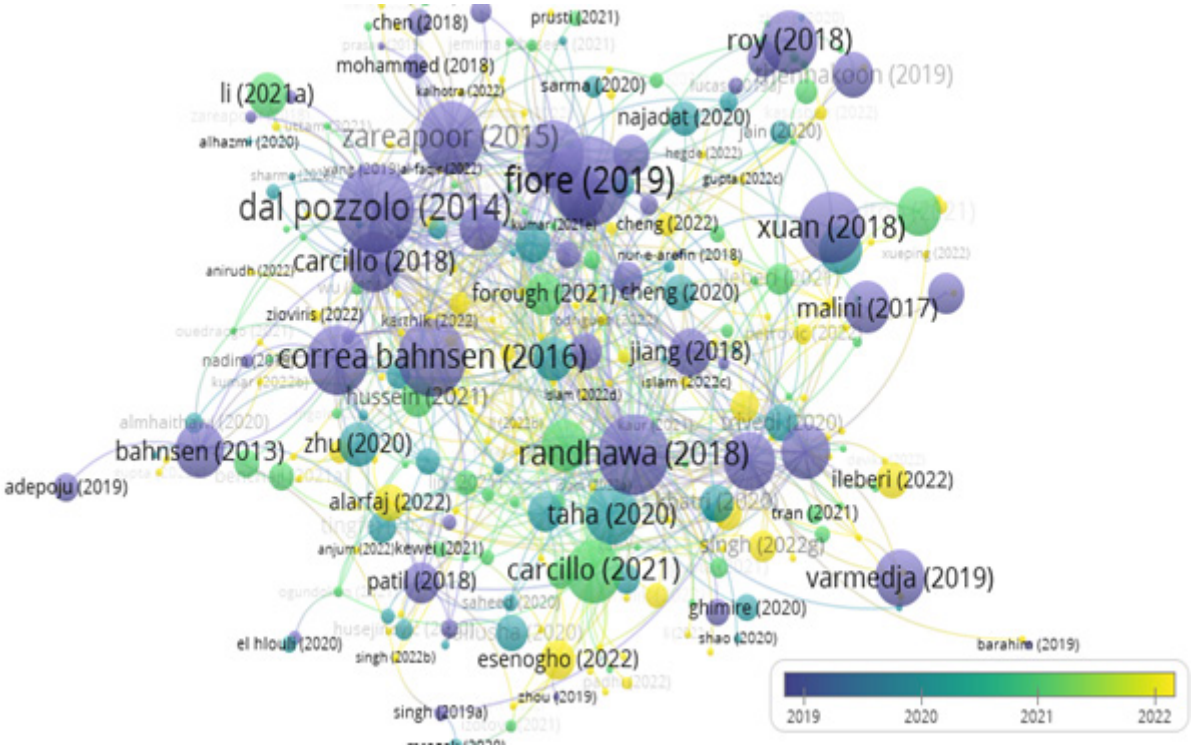

Figure 3: Citation map of documents published on CCF detection.

Table 1: Most relevant and local cited sources.

Most relevant Sources		Most Local Cited Sources		Source Impact		
Sources	**No.**	**Sources**	**No.**	**Sources**	***h*-index**	***g*-index**
Advances In Intelligent Systems and Computing	21	IEEE Access	984	IEEE Access	9	10
ACM International Conference Proceeding	20	Expert Systems with Applications	718	Advances in Intelligent Systems and Computing	6	9
Lecture notes in Networks and Systems	19	Information Sciences	568	Procedia Computer Science	6	7
Lecture notes in Electrical Engineering	15	Procedia Computer Science	459	Expert Systems with Applications	5	5
IEEE Access	10	International Journal of Advanced Computer Science and Applications	394	ACM International Conference Proceeding Series	4	12
Lecture Notes in Computer Science	8	ACM International Conference Proceeding Series	150	International Conference on Communications and Electronic Systems	3	3
Communications in Computer and Information Science	7	Advances in Intelligent Systems and Computing	97	Arabian Journal for Sciences and Engineering	3	3
Procedia Computer Science	7	Lecture Notes in Electrical Engineering	80	Information Sciences	3	3
AIP Conference Proceedings	5	Wireless Communications and Mobile Computing	17	International Journal of Interactive Mobile Technologies	3	3
Expert Systems with Applications	5	Journal of Physics Conference Series	13	Journal of Big Data	3	3

nodes in different sizes and color themes. The size of the node describes the number of times an author has been cited.

Subsequently, the number of documents published by the authors and their citation metrics identify the most active researchers in the OCCF detection domain, as shown in Table 2 Lui. G and Kumar. A recorded the highest number of documents. However, the number of citations for Caelen. O was the highest followed by Bontempi. G and Le Borgne. Y respectively. The average citation per document provides more information about the impact of publications and by extension, the author. In other words, a good-quality article would receive a higher number of citations. It can be observed that Caelen. O (publications: 9, citations 558) and Bontempi. G, (publications:8, citations 487) had the highest average citation per publication. This is suggestive of the fact that the documents of the two authors are more impactful compared to the other publications in this research domain. Kumar. A and Liu. G produced the highest number of publications as previously mentioned with 11 apiece. Interestingly the average citation score and link strength of Kumar. A is very low. Both authors' link strength and citations are lower than that of Caelen. O, Bontempi. G and Le Borgne. Y indicates that the latter three researchers exhibit good networking and collaborating numbers.

Key concepts and trending topics in literature on online credit card fraud detection

To address research question 3, a description of the interaction of keywords, authors, and sources is given in Figure 6, a three-field plot. A field plot is an interactive approach to describe information, it is based on the concept of Sankey diagrams, which are mainly used to pictorially depict the movement of energy or materials in several processes and networks.[45] This was designed in Biblioshiny.[37] The rectangle shapes on the plot represent elements of keywords, authors, and sources where the size of the shape is proportional to the value of the element in terms of frequency. For example, the keyword fraud detection is the most frequent keyword in this analysis, followed by machine learning and then advances in intelligent systems and computing. In the middle of the plot, "DE" depicts the 15 most frequent keywords, which is also the topic of the publications. On the left-hand side, "AU" depicts the authors related to these topics, and on the right side "SO" represents the publication sources. This information helps identify which authors are publishing frequently and on what topics. SO also identifies which are the more consistent publishers and in which topic for example, for keyword smote, Kumar A, has published the most and the most active publishers are Information Sciences, Communications

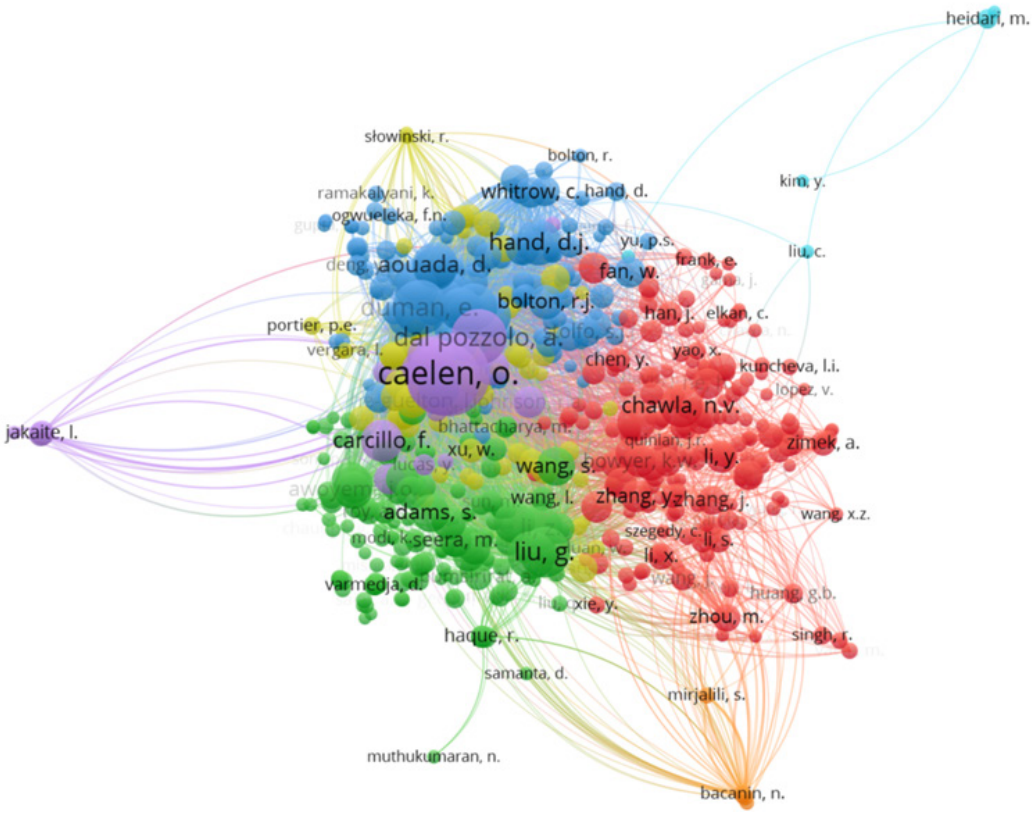

Figure 4: Co-citation map of authors in CCF detection.

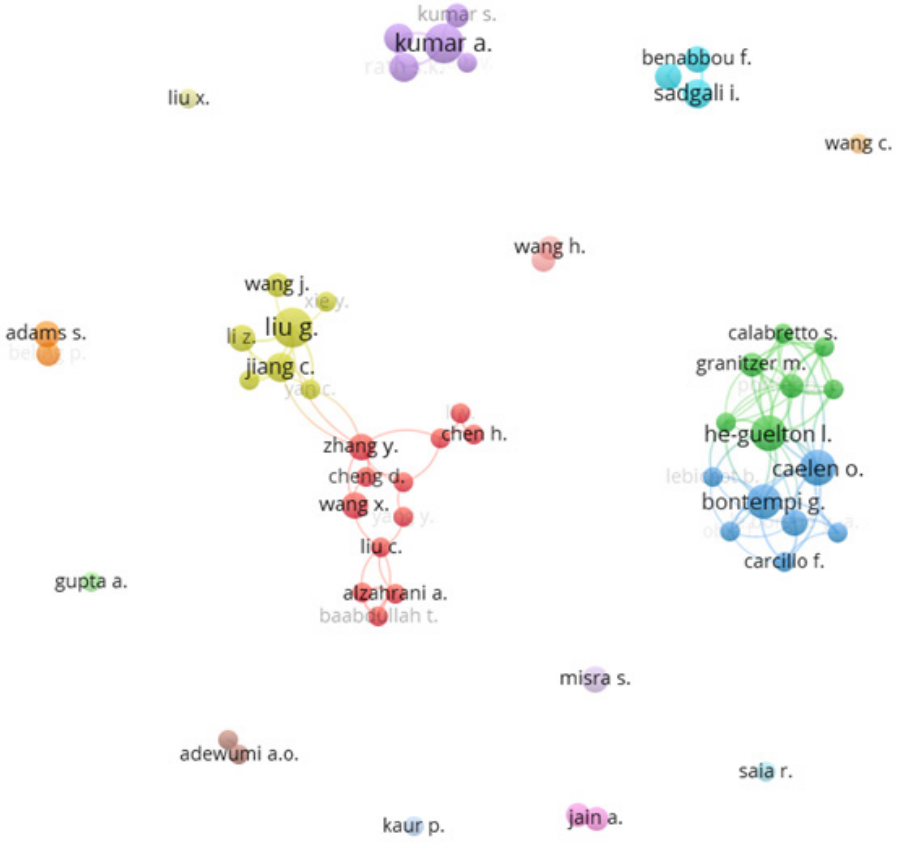

Figure 5: Authors and co-citation relationships on OCCF detection.

Table 2: Authors with more than four publications.

ID	Author	Documents	Citations	Average Citations	Total Link Strength
1	Kumar A.	11	40	3.64	7
2	Liu G.	11	309	28.09	32
3	Caelen O.	9	558	62.00	48
4	He-guelton l.	9	89	9.89	18
5	Bontempi G.	8	487	60.88	48
6	Jiang C.	6	260	43.33	29
7	Prusti D.	6	25	4.17	0
8	Rath S. K.	6	25	4.17	0
9	Sadgali I.	6	22	3.67	12
10	Adams S.	5	168	33.60	3
11	Benabbou F.	5	14	2.80	6
12	Le Borgne Y.	5	418	83.60	45
13	Li Z.	5	188	37.60	16
14	Misra S.	5	2	0.40	16
15	Sael N.	5	14	2.80	6
16	Wang X.	5	24	4.80	9
17	Zhang Y.	5	46	9.20	7
18	Beling P.	4	159	39.75	3
19	Granitzer M.	4	70	17.50	4
20	Jain A.	4	29	7.25	6
21	Kumar S.	4	2	0.50	2
22	Portier P.-E.	4	70	17.50	4
23	Singh A.	4	34	8.50	0
24	Wang H.	4	33	8.25	1
25	Wang J.	4	6	1.50	8
26	Zhang X.	4	66	16.50	8

in Computer and Information Sciences, and Lecture Notes in Electrical Engineering.

The plot depicting the titles, authors, and keywords in a three-field format illustrates the relationship between publications and the important concepts related to OCCF detection. The process of source analysis involved selecting the most pertinent and regionally acknowledged sources for the compiled dataset. The assessment of productivity in the realm of OCCF detection is based on the identification of the most pertinent and widely referenced sources, which provide information on the authors and titles cited in the references list of each document. The Journal for Advances in Intelligent Systems and Computing emerged as the primary source for obtaining the most pertinent information on fraud detection and credit card fraud detection, encompassing a total of 22 documents. Subsequently, lecture notes in networks and systems and lecture notes in electrical engineering exhibited significant relevance in this domain, comprising 19 and 15 documents respectively. Further, the 15 most commonly occurring keywords within the domain are also depicted in Figure 6. Notably, the topic of fraud detection emerges as the dominant theme among authors and is present across all the journals. The keyword in question is closely associated with the keywords machine learning and credit card fraud, in that order.

Subsequently, to determine the key concepts and trending topics in OCCF detection, co-occurrence analysis was performed using all keywords and a combination of all authors' keywords, resulting in a map depicted in Figure 7. This map shows the most frequent words according to the size of the cluster. This network analysis is important to identify key concepts that have been explored in OCCF detection and how these concepts relate to each other. The minimum number of occurrences of keywords chosen for this analysis is 5, and 209 out of a total of 2193 met this threshold. The keywords in the same colors represent the same school of thought. For the same color, these keywords have been co-authored by

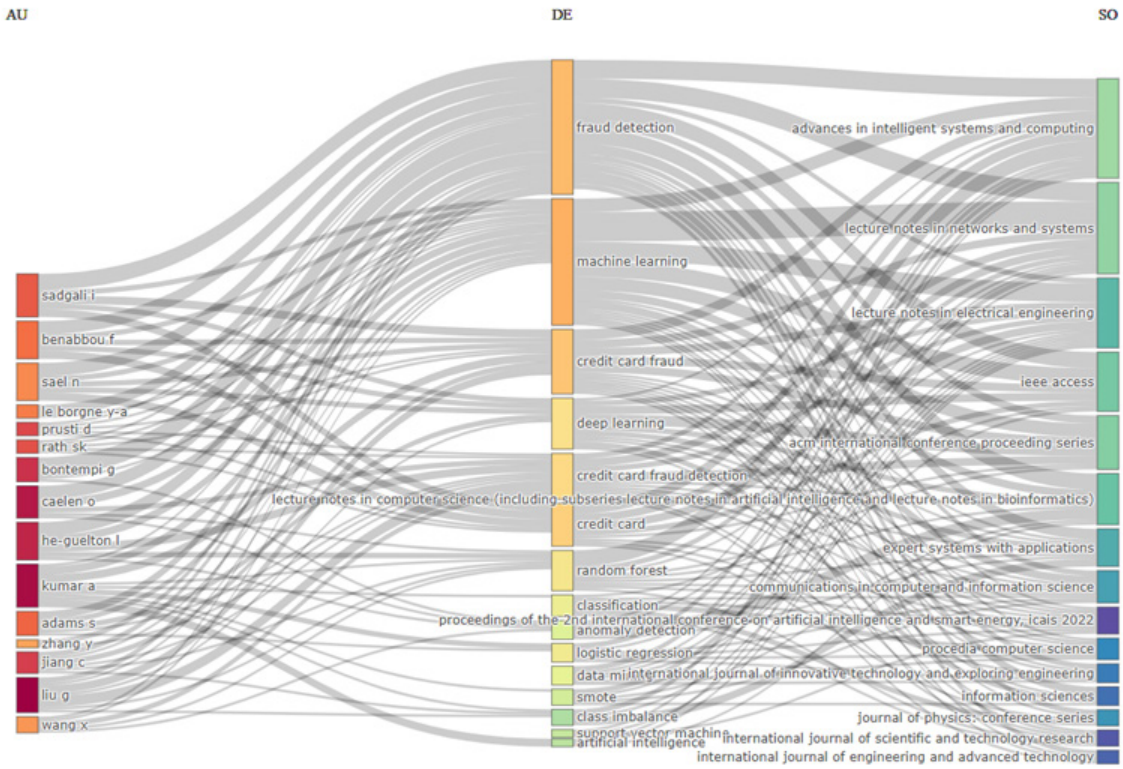

Figure 6: Three-field plot on keywords, authors, and sources.

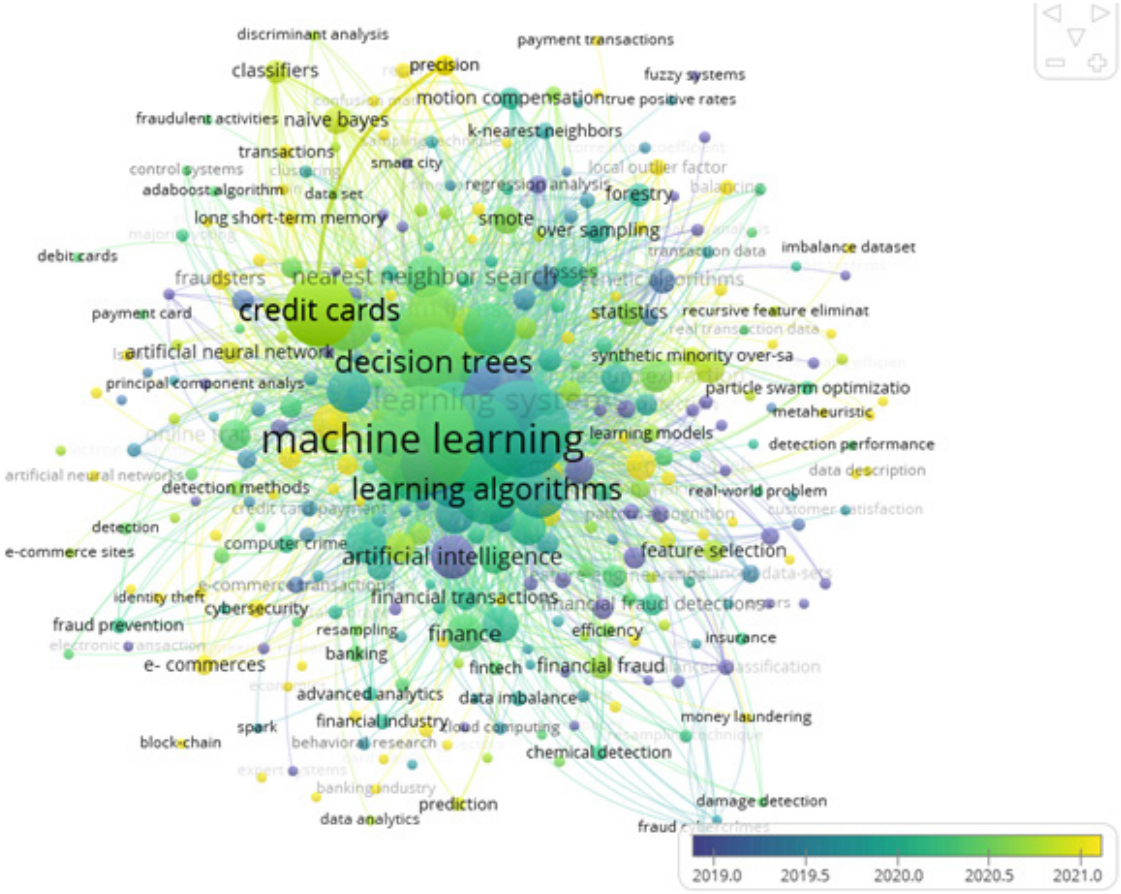

Figure 7: Keyword map in chronological order of occurrence.

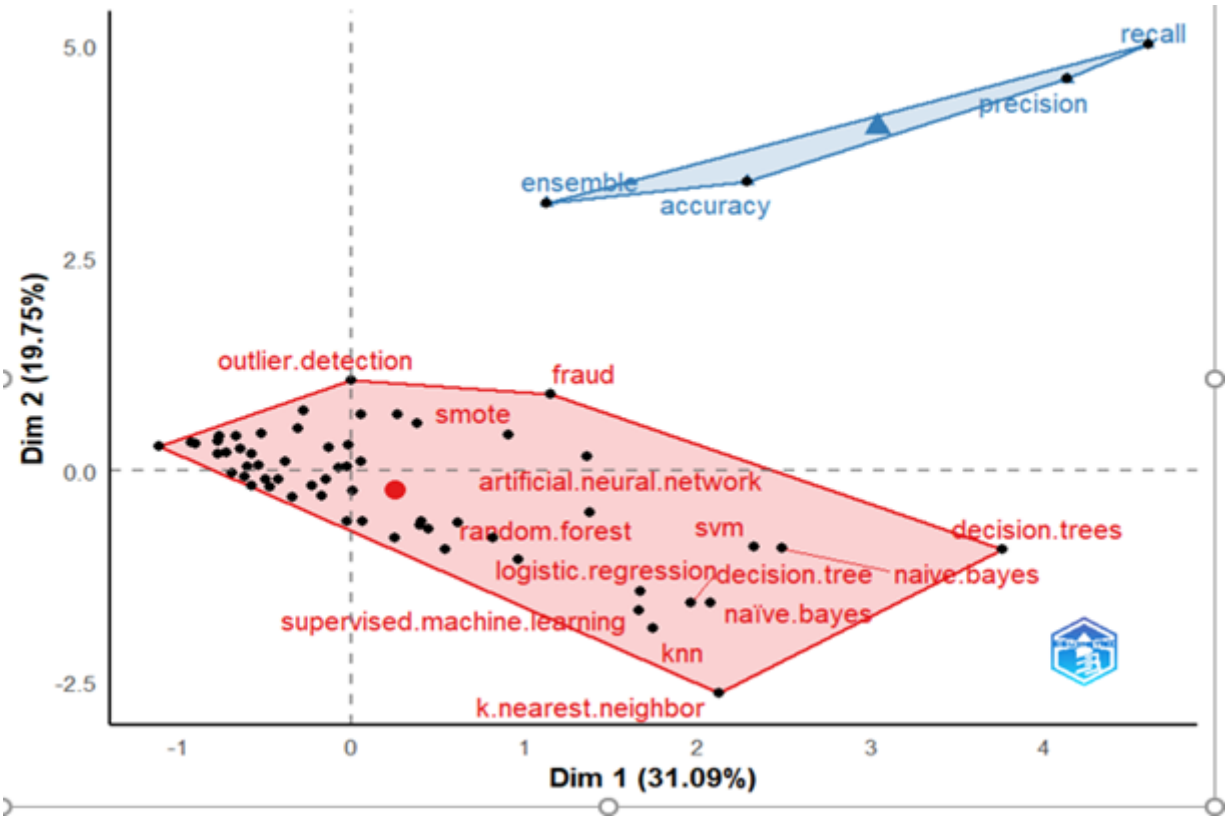

Figure 8: Conceptual structure map.

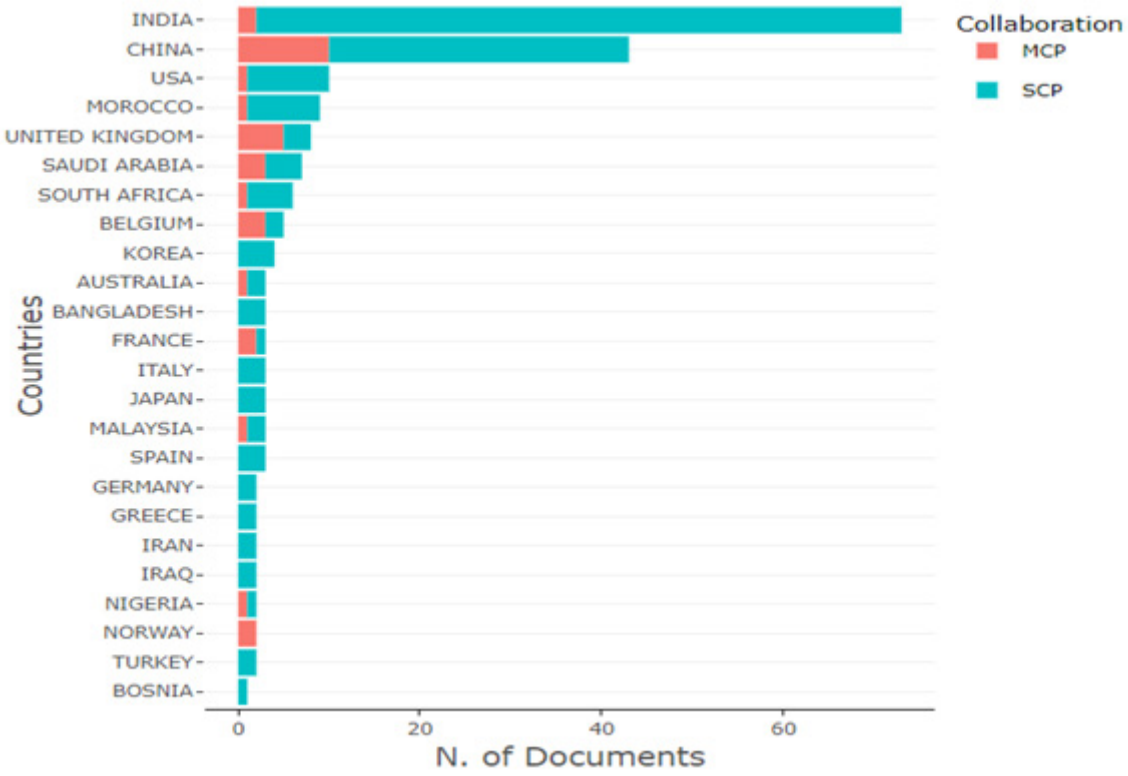

Figure 9: Most productive countries.

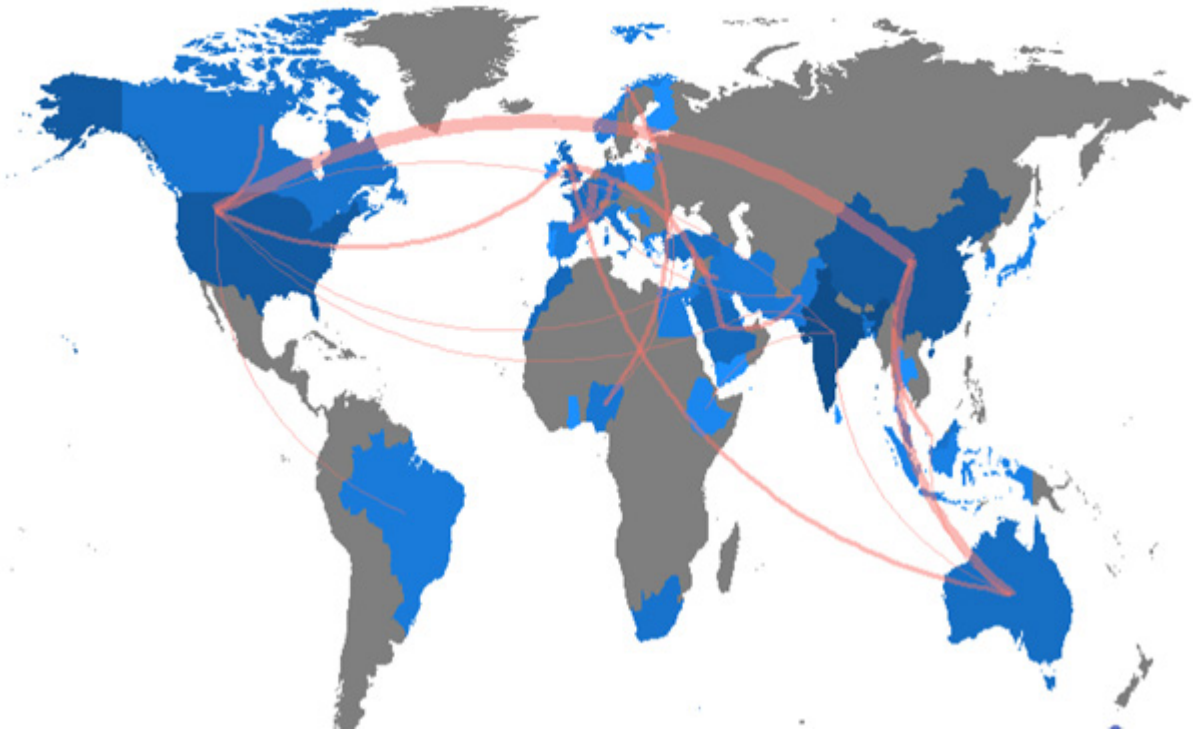

Figure 10: Collaboration world map.

the authors in the links for that theme. Consequently, machine learning is the most frequently repeated occurring keyword on the topic of OCCF detection (172) followed by learning algorithms (150) and credit cards (86).

To further demonstrate key concepts in the study, Figure 7 depicts the evolution of concepts over time with the dark purple color relating to historical keywords and the yellow color representing the most recent key terms. Initially, there was more focus on data mining techniques in this domain as depicted by the purple color in the nodes. From the period mid-2019 and mid-2020, the focus gradually shifted to ML and feature selection in OCCF detection. Ultimately, as shown by the yellow-green color on the map, studies on OCCF detection are now more concentrated on EL.

This concentration on EL is also depicted in Figure 8, a conceptual structure map through the execution of Multiple Correspondence Analysis (MCA) of authors' keywords. This could be because of the benefits inherent in combining more than one approach to address the OCCF detection challenge making EL approaches the most trending area. Applying MCA to all keywords in the publications allows Figure 8 to show the keywords' conceptual structure according to the publications. As shown, extensive data is compressed with several variables producing a low-dimensional space of an intuitive two-dimensional graph shown as plane distance which identifies any resemblance among keywords. OCCF is a form of fraud; therefore, the keyword fraud has received much attention evidenced by its position at the center point. Further, to other points subsequent distribution and positions along the dimensions, keywords that are similar in distribution are represented in the same cluster.[37] Figure 8 shows two clusters, red and blue. Red is the most significant cluster consisting of 14 keywords which are outlier detection, fraud, smote, ANN, RF, SVM, decision tree, LR, KNN, naïve Bayes, supervised machine learning, random forest, and k nearest neighbor. Blue cluster comprises 4 keywords which are ensemble, accuracy, precision, and recall.

Distribution and nature of collaboration at the country level

There are 642 documents issued by 63 different nations. India produced the most publications (198 documents, or 30.6% of the total), followed by China (71 publications, or 10.97% of the total). The United States is ranked third with 67 publications (10.36%) while Morocco and the United Kingdom are tied for fourth with 22 publications (3.4% each). In the subsequent years, eleven countries produced between 10 and 19 documents, while ten countries published between 5 and 9 documents. In addition, 17 countries have contributed between 2 and 4 documents, while 20 have published a single document. Adding the contributions from each country yields a total of 642 publications, which is greater than 524. This indicates that there has been an abundance of international collaboration. Table 3 provides statistics for a selection of the top 46 countries. Table 3 reveals that just 19 nations had a nominal GDP ranking of 20 or lower. This demonstrates that the more economically developed nations are likewise concerned about the persistence of OCCF issues

Table 3: Countries that published more than 4 documents.

ID	Country	Documents	Percentage %	Citation	Average citation per document	Nominal GDP Rank*	Total Link Strength
1	India	198	30.60	1029	5.20	5	20
2	China	71	10.97	963	13.56	2	31
3	United States	67	10.36	1419	21.18	1	40
4	Morocco	22	3.40	108	4.91	60	1
5	United Kingdom	22	3.40	466	21.18	6	31
6	Saudi Arabia	19	2.94	197	10.37	18	22
8	France	16	2.47	246	15.38	7	17
9	Australia	13	2.01	224	17.23	14	21
11	Turkey	12	1.85	12	1.00	20	3
12	Belgium	12	1.85	531	44.25	26	12
13	Germany	12	1.85	83	6.92	4	7
14	Bangladesh	11	1.70	29	2.64	35	1
15	Malaysia	10	1.55	195	19.50	36	11
16	Canada	10	1.55	131	13.10	8	10
17	Nigeria	10	1.55	226	22.60	31	11
18	Taiwan	9	1.39	184	20.44	21	5
19	South Africa	9	1.39	107	11.89	39	2
20	Vietnam	9	1.39	54	6.00	38	8
21	Italy	7	1.08	328	46.86	10	3
22	Iran	7	1.08	224	32.00	11	3
23	Iraq	7	1.08	35	5.00	48	5
24	Spain	7	1.08	184	26.29	16	1
25	Brazil	6	0.93	48	8.00	12	4
26	Egypt	5	0.77	38	7.60	33	1
27	Norway	5	0.77	15	3.00	30	11
28	Jordan	4	0.62	28	7.00	92	0
29	Pakistan	4	0.62	49	12.25	42	8
30	South Korea	4	0.62	77	19.25	13	0
31	Japan	3	0.46	6	2.00	92	0
32	Ethiopia	3	0.46	46	15.33	67	7
33	Indonesia	3	0.46	13	4.33	17	2
34	Serbia	3	0.46	52	17.33	87	0
35	Bahrain	2	0.31	2	1.00	94	2
36	Greece	2	0.31	0	0.00	54	0
37	Israel	2	0.31	20	10.00	28	2
38	Kuwait	2	0.31	16	8.00	59	3
39	Lithuania	2	0.31	2	1.00	85	4
40	Luxembourg	2	0.31	271	135.50	70	0
41	Russian Federation	2	0.31	7	3.50	9	0
42	United Arab Emirates	2	0.31	16	8.00	32	1
43	Ireland	2	0.31	2	1.00	29	0

ID	Country	Documents	Percentage %	Citation	Average citation per document	Nominal GDP Rank*	Total Link Strength
44	Poland	2	0.31	29	14.50	23	5
45	Hong Hong	1	0.15	0	0.00	43	2
46	Portugal	1	0.15	145	145.00	50	1

and are researching all research-based options for combating this offense. The United States of America has received the most citations per document, with an average of 21.18 citations per document from 67 publications published in the country. Even though India produced the most documents, the average number of citations per document for this country is 5, 20, resulting in a total of 1029 citations. China, Belgium, the United Kingdom, Italy, Luxembourg, France, Nigeria, Iran, and Australia are among the countries with a relatively high number of citations recorded at 200 or more. It is vital to note that India and China rank top and second, respectively, in terms of the number of published documents. However, China is ranked 24th and India is ranked 42nd for citations.

The total strength link strength provides an estimation of the collaborative research of one country with the other countries. Further analysis using VosViewer was conducted to identify the collaboration between different geographical regions and interesting results were produced. All the continents have some publications in this domain with India and China the *busiest* countries as shown in Figure 9.

Additional information was extracted on the geographical analysis using Biblioshiny to support results obtained through VOSViewer as well as to reduce bias. The results were similar as shown in Figure 10, showing a commendable social structure of scholarly collaboration among three countries in this domain. However, there is a strong collaboration link between North America, Asia, and Europe. South America and Africa have no strong links with other regions in terms of collaborations. There is no link between the two regions indicating a lack of collaboration. More so, from the world map in Figure 10, we deduce that most of the collaboration efforts are initiated in the United States of America and that, the region has collaborated with all regions except Africa.

CONCLUSION

A bibliometric analysis of OCCF detection trends was presented. The citation and publication patterns of literature from 2013 to 2022 were analyzed. The structure of the knowledge base and the most impactful authors in this domain were also presented. Further, using both Biblioshiny and VOSViewer, the nature of collaboration between countries and authors and the key themes, including trends to guide future research, were also outlined.

The number of articles produced from year to year is considered a reliable indicator of the research trend in any research field. Research in the OCCF domain started to gather momentum post 2017 reflecting that this area is still problematic because the numbers continue to surge, recorded at 144 publications in 2021 alone.

An analysis of the authors' and co-authors' relationships was also conducted to identify the major research groups that are working in a particular field. The study indicates that there are four major groups, considering at least three authors in a group. The research group of Zhang. Y., and Liu, G. are highly connected and form the largest cluster of 19 researchers. Further, Caelene, O. has been identified as one of the most productive authors with nine publications and total link strength of 32. All of the authors in the study are represented by a network of nodes of varying sizes and color themes. The size of the node describes the number of times an author has been cited. The number of documents published by the authors, as well as their citation metrics, identify the most active researchers in the domain of OCCF detection. Subsequently, Caelen. O. (publications: 9, citations: 558) and Bontempi. G. (publications: 8, citations: 487) had the highest average citation per publication.

The study further described the interaction of keywords, authors, and sources in the three-field plot. This plot is based on the concept of Sankey diagrams and identifies the 15 most frequent keywords, which are also, interestingly, topics of publications in this analysis. However, the conceptual structure of the keywords has been depicted through the application of MCA to all the keywords, resulting in the word ensemble as the most trending research technique in OCCF detection even though the word fraud has attracted more attention, as evidenced by its position at the center point. This indicates that the trends in OCCF detection are moving towards ensemble learning techniques.

The analysis identified 642 documents issued by 63 different nations. India produced the most publications (198 publications, or 30.6% of the total), followed by China (71 publications, or 10.97%). The United States is ranked third with 67 publications (10.36%), while Morocco and the United Kingdom are tied for fourth with 22 publications (3.4%) each.

Therefore, there is a sharp upward trajectory in the total number of publications in OCCF detection starting from 2014. In this

trajectory, Asia is the most productive region in the scholarly publication domain, with India taking the lead and closely followed by China. However, in terms of collaboration, most collaborations are initiated in the United States, and there is a link amongst all the regions other than Africa and South America, an indication of strong collaboration initiatives amongst those regions.

We also discovered that, through the mapping of knowledge, five distinct schools of thought or themes were discovered in scholarly publications. These schools of thought are imbalanced data, anomaly detection, machine learning, decision trees, and ensemble learning. The themes are a general group of thoughts around the area of OCCF detection space. Publication documents in this study are clustered according to these themes. Therefore, current and future researchers could benefit by comprehending and acknowledging the existence of these five schools of thought in OCCF detection.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Dhankhad S, Mohammed E, Far B, editors. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. 2018 IEEE international conference on information reuse and integration (IRI); 2018: IEEE.
2. Zadafiya N, Karasariya J, Kanani P, Nayak A, editors. Detecting Credit Card Frauds Using Isolation Forest And Local Outlier Factor-Analytical Insights. 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT); 2022: IEEE.
3. Yuan M, A editor Transformer-based Model Integrated with Feature Selection for Credit Card Fraud Detection. 7th International Conference on Machine Learning Technologies (ICMLT); 2022.
4. Nguyen VB, Dastidar KG, Granitzer M, Siblini W. The Importance of Future Information in Credit Card Fraud Detection. arXiv preprint arXiv:220405265. 2022.
5. Thennakoon A, Bhagyani C, Premadasa S, Mihiranga S, Kuruwitaarachchi N, editors. Real-time credit card fraud detection using machine learning. 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence); 2019: IEEE.
6. Sadgali I, Nawal S, Benabbou F, editors. Fraud detection in credit card transaction using machine learning techniques. 1st International Conference on Smart Systems and Data Science (ICSSD); 2019: IEEE.
7. Yang Y, Yu Y, Li T, editors. Deep Learning Techniques for Financial Fraud Detection. 14th International Conference on Computer Research and Development (ICCRD); 2022: IEEE.
8. Singh A, Ranjan RK, Tiwari A. Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms. Journal of Experimental & Theoretical Artificial Intelligence. 2022;34(4):571-98.
9. Maurya A, Kumar A, editors. Credit card fraud detection system using machine learning technique. IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom); 2022: IEEE.
10. Esna-Ashari M, Khamesian F, Khanizadeh F. Using local outlier factor to detect fraudulent claims in auto insurance. Journal of Mathematics and Modeling in Finance. 2022;2(1):167-82.
11. Showalter S, Wu Z. Minimizing the Societal Cost of Credit Card Fraud with Limited and Imbalanced Data. arXiv preprint arXiv:190901486. 2019.
12. Stellwall M. Learn Data Science [Internet]: Oracle AI & Data Science Blog. 2019. Available from: https://blogs.oracle.com/ai-and-datascience/post/overview-of-traditional-machine-learning-techniques.
13. Padhi B, Chakravarty S, Biswal B. Anonymized credit card transaction using machine learning techniques. Advances in Intelligent Computing and Communication: Springer; 2020. p. 413-23.
14. Ata O, Hazim L. Comparative analysis of different distributions dataset by using data mining techniques on credit card fraud detection. Tehnički vjesnik. 2020;27(2):618-26.
15. Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences. 2019;479:448-55.
16. Zhang Z, Chen L, Liu Q, Wang P. A fraud detection method for low-frequency transaction. IEEE Access. 2020;8:25210-20.
17. Gupta K, Singh K, Singh GV, Hassan M, Sharma U, editors. Machine Learning based Credit Card Fraud Detection-A Review. International Conference on Applied Artificial Intelligence and Computing (ICAAIC); 2022: IEEE.
18. Abakarim Y, Lahby M, Attioui A, editors. An efficient real time model for credit card fraud detection based on deep learning. Proceedings of the 12th international conference on intelligent systems: theories and applications; 2018.
19. Alarfaj FK, Malik I, Khan HU, Almusallam N, Ramzan M, Ahmed M. Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms. IEEE Access. 2022;10:39700-15.
20. Ali I, Aurangzeb K, Awais M, Aslam S, editors. An Efficient Credit Card Fraud Detection System using Deep-learning based Approaches. IEEE 23rd International Multitopic Conference (INMIC); 2020: IEEE.
21. Divya P, Palanivel Rajan D, Selva Kumar N. Analysis of machine and deep learning approaches for credit card fraud detection. ICCCE 2020: Springer; 2021. p. 243-54.
22. Prabha N, Manimekalai S, editors. Imbalanced data Classification in Credit Card Fraudulent Activities Detection using Multi-Class Neural Network. Second International Conference on Artificial Intelligence and Smart Energy (ICAIS); 2022: IEEE.
23. Rao GM, Srinivas K. RNN-BD: an approach for fraud visualisation and detection using deep learning. International Journal of Computational Science and Engineering. 2022;25(2):166-73.
24. Shaji A, Binu S, Nair AM, George J, editors. Fraud detection in credit card transaction using ann and svm. International Conference on Ubiquitous Communications and Network Computing; 2021: Springer.
25. Benchaji I, Douzi S, El Ouahidi B. Credit card fraud detection model based on LSTM recurrent neural networks. Journal of Advances in Information Technology. 2021;12(2).
26. Laveti RN, Mane AA, Pal SN, editors. Dynamic Stacked Ensemble with Entropy based Undersampling for the Detection of Fraudulent Transactions. 6th International Conference for Convergence in Technology (I2CT); 2021: IEEE.
27. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. Frontiers of Computer Science. 2020;14(2):241-58.
28. Forough J, Momtazi S. Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing. 2021;99:106883.
29. González S, García S, Del Ser J, Rokach L, Herrera F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion. 2020;64:205-37.
30. Karthik V, Mishra A, Reddy US. Credit Card Fraud Detection by Modelling Behaviour Pattern using Hybrid Ensemble Model. Arabian Journal for Science and Engineering. 2022;47(2):1987-97.
31. Tomar P, Shrivastava S, Thakar U, editors. Ensemble Learning based Credit Card Fraud Detection System. 5th Conference on Information and Communication Technology (CICT); 2021: IEEE.
32. Trisanto D, Rismawati N, Mulya MF, Kurniadi FI. Modified Focal Loss in Imbalanced XGBoost for Credit Card Fraud Detection. Int J Intell Eng Syst. 2021;14:350-8.
33. Bayram B, Köroğlu B, Gönen M, editors. Improving fraud detection and concept drift adaptation in credit card transactions using incremental gradient boosting trees. 19th IEEE International Conference on Machine Learning and Applications (ICMLA); 2020: IEEE.
34. Feng H, editor Ensemble learning in credit card fraud detection using boosting methods. 2nd International Conference on Computing and Data Science (CDS); 2021: IEEE.
35. Al Rubaie EMH. Improvement in credit card fraud detection using ensemble classification technique and user data. International Journal of Nonlinear Analysis and Applications. 2021;12(2):1255-65.
36. Shukla S, Rakesh D, editors. Dynamic ensemble based feature selection model for credit card fraud detection. IEEE 17th India Council International Conference (INDICON); 2020: IEEE.
37. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of informetrics. 2017;11(4):959-75.
38. Zhang L, Ling J, Lin M. Artificial intelligence in renewable energy: A comprehensive bibliometric analysis. Energy Reports. 2022;8:14072-88.
39. Chen Y, Lin M, Zhuang D. Wastewater treatment and emerging contaminants: Bibliometric analysis. Chemosphere. 2022;297:133932.
40. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial intelligence in health care: bibliometric analysis. Journal of Medical Internet Research. 2020;22(7):e18228.

41. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology. 2012;63(8):1609-30.
42. Hallinger P, Kovačević J. A bibliometric review of research on educational administration: Science mapping the literature, 1960 to 2018. Review of Educational Research. 2019;89(3):335-69.
43. Crocetti E. Systematic reviews with meta-analysis: Why, when, and how? Emerging Adulthood. 2016;4(1):3-18.
44. Verborgh R, De Wilde M. Using OpenRefine: Packt Publishing Ltd; 2013.
45. Riehmann P, Hanfler M, Froehlich B, editors. Interactive sankey diagrams. IEEE Symposium on Information Visualization, INFOVIS 2005: IEEE.

Cite this article: Hove D, Olugbara O, Singh A. Bibliometric Analysis of Recent Trends in Machine Learning for Online Credit Card Fraud Detection. J Scientometric Res. 2024;13(1):43-57.