IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE EDUCATION SOCIETY SECTION

Received 25 April 2023, accepted 15 May 2023, date of publication 25 May 2023, date of current version 6 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280075

RESEARCH ARTICLE

Analysis of Attrition Studies Within the Computer Sciences

GEORGE OBAIDO[1], (Member, IEEE), FRIDAY JOSEPH AGBO[2,3], CHRISTINE ALVARADO[4], AND SOLOMON SUNDAY OYELERE[5]

[1]Center for Human-Compatible Artificial Intelligence, Berkeley Institute for Data Science, University of California at Berkeley, Berkeley, CA 94720, USA
[2]School of Computing and Information Sciences, Willamette University, Salem, OR 97301, USA
[3]School of Computing, University of Eastern Finland, 80130 Joensuu, Finland
[4]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093, USA
[5]Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, 93187 Skellefteå, Sweden

Corresponding author: George Obaido (rabeshi.george@gmail.com)

ABSTRACT Student attrition is a long-standing problem in Computer Science (CS), as in many other disciplines, and it has gained momentum in the academic sphere. This study employs *bibliometric analysis* to shed light on the research stream of student attrition within CS. Bibliometric analysis is a popular technique for evaluating published scientific articles when empirical contributions are producing voluminous research streams. We collected 1310 articles from the Web of Science and Scopus databases, published over a period of 22 years from 2000 to 2022, to analyze the most relevant publication venues in the study of attrition in CS. Further analysis revealed the most cited institutions, countries, key themes, and other conceptual information. Keywords, such as ''retention,'' ''computer science education,'' ''gender,'' ''introductory programming,'' and ''student success'' emerged as dominant themes in attrition studies. As researchers work intensively to reduce attrition within CS, these thematic areas may continue to shape the future direction of attrition studies. Our study provides a comprehensive overview of research hotspots, thematic areas, and future directions for attrition studies in CS. This outcome could be valuable for young and emerging scholars who are starting their careers and looking to identify research hotspots in this field of interest.

INDEX TERMS Bibliometric analysis, scientific mapping, computer science education, student attrition, retention, at-risk students.

I. INTRODUCTION

Within Computer Science (CS), student attrition continues to be a long-standing problem and remains a significant challenge for the major [1], [2], [3], [4]. Student attrition affects institutions of higher learning globally and can be damaging to their reputation [5], [6], [7]. Several significant reasons have emerged for the high rate of attrition within CS: poor project management skills, lack of understanding of the material, not identifying with the career path, cultural issues, lack of assistance and feedback, poorly designed courses, personal problems, and more [8], [9], [10]. As a result, CS education researchers have extensively studied this problem, using various tools and techniques to make early interventions before students drop out. However, academic publications on this topic are growing exponentially.

Researchers have applied quantitative, qualitative, or mixed methods to understand attrition [11], [12], [13], [14]. For example, Kinnunen and Malmi [8] performed a quantitative survey to understand root causes behind attrition in CS. Sharmin [15] examined the potential of teaching creativity as a skill for retaining undergraduate students in CS. The study explored Keller's Attention, Relevance, Confidence, Satisfaction (ARCS) motivational model [16], which advocates for open-ended assignments, collaborative learning, and other strategies to reduce attrition. A recent review advocated for diversity in computer science programs as a means of reducing attrition [1]. The study noted that tackling attrition issues is necessary to increase diversity. Despite the impact of these studies, bibliometric analysis can provide a

The associate editor coordinating the review of this manuscript and approving it for publication was John Mitchell.

systematic and reproducible background of this topic. Given the overwhelming amount of literature on this topic and various intervention strategies and techniques, bibliometric analysis is useful in presenting findings related to the theme of the study, trends over time, the most prolific scholars and institutions, and the perspective of different countries [17], [18], [19], [20].

According to Godin [20], bibliometric analysis is a subfield that measures the output side of science. Bibliometric methods have been used in various forms, and are often used interchangeably with broader terms, such as *scientometrics* or *infometrics* [21], [22], [23], [24]. Bibliometric analysis is becoming a popular method for providing quantitative descriptions of published articles [22], [23], [25]. This method has found applications in several areas, such as CS education [18], [26], medical [27], [28] and engineering [29], and international entrepreneurship [30]. It uses statistical and geometrical methods in diverse analysis and mapping tools to measure knowledge domains. These tools are primarily used on bibliographic databases, such as Google Scholar (GS), Microsoft Academic (MA), Web of Science (WoS), and Scopus [17], [31]. Additionally, these tools require little or no programming skills to use [17]. Most of these tools exist to analyze the impact of a scientific topic and its structure, while some are no longer maintained [17], [32], [33], [34]. Examples include `Bibliometrix R-package` [17], `SciMAT` [32], and `CitNetExplorer` [35], `VOSviewer` [33], [34], [36].

Our analysis shows that publications on CS attrition are growing exponentially. To the best of our knowledge, no study has examined bibliometric analysis to analyze trends of publications and research constructs in this area, especially using the Scopus and Web of Science databases. Our study provides a blueprint for young and emerging researchers who want to focus their research on attrition in CS studies. By analyzing publication trends, collaborative networks, and popular keywords in this domain, our study can provide useful information to these scholars. Therefore, our study aims to answer the following questions:

- RQ1 What is the number of publications on this topic over the past twenty-two years?
- RQ2 Which are the top publication venues on this topic?
- RQ3 How has the keyword usage grown over these years?
- RQ4 What are the commonly used themes in studies on student attrition within computer science?
- RQ5 What are the commonly used keywords by authors in this research field during this period?

These questions aim to provide insights to deepen our understanding of attrition research within the computer science discipline, identify commonly used keywords in this field, and highlight gaps to be addressed in future studies by young and emerging scholars. The remainder of this study is organized as follows: Section II presents the methods used in this study, Section III presents the results of the bibliometric analysis, Section IV discusses the results, and Section V concludes the study and suggests areas for future research.

II. METHODOLOGY

Bibliometric analysis was the chosen technique for this study, and the recommended workflow for scientific mapping process by Aria and Cuccurullo [17] and Agbo et al. [18] was employed. The `Bibliometrix` package and the `biblioshiny` function in the R programming language were used for the quantitative analysis tasks, and the VOSviewer software was used to create and visualize the bibliometric structures. The methodology workflow is presented in Figure 1.

A. DATA COLLECTION AND EXTRACTION

Table 1 presents the selection criteria for articles contained in this study. We extracted data from two popular databases: Clarivate Analytics Web of Science (WoS) and Scopus, which contained millions of records. We limited our scope to articles published between January 1, 2000 through December 29, 2022, and we did not restrict the document type to any specific type. We used Boolean queries with keyword terms related to computer education, computer science, and computer engineering, as well as retention, attrition, student retention, drop-out prevention, student success, support strategies, and student integration, as shown in Table 1. Finally, we restricted our set to include only articles written in English. Our final set comprised 1310 articles including journals, conference proceedings, book chapters, and other electronic sources authored by 3143 individuals. A summary of the search results is presented in Table 2.

B. DATA ANALYSIS

The data analysis is divided into three phases. The *first phase*, descriptive analysis, involves summarizing and displaying the bibliographic data, which typically includes information about the articles, such as the author, publication date, journal, and keywords. The purpose of this phase is to gain a general understanding of the data and to identify any trends or patterns that may emerge. The *second phase*, network analysis, involves creating co-occurrence and co-citation networks to explore the relationships between the different articles and keywords. Co-occurrence networks display the frequency of appearance of different keywords or terms in the same article, while co-citation networks show the frequency of appearance of two articles in the same study. These networks help identify clusters of related articles and keywords, which can be useful in identifying research trends and areas of interest. The *third phase*, normalization, involves generating a similarity measure of the attrition dataset [17].

We used two measures to calculate the similarity between different sets of articles: Jaccard's coefficient and Salton's cosine. Jaccard's coefficient and Salton's cosine are two similarity measures commonly used in bibliometric analysis to assess the similarity between two sets of items, such as articles or keywords [37], [38], [39]. Jaccard's coefficient

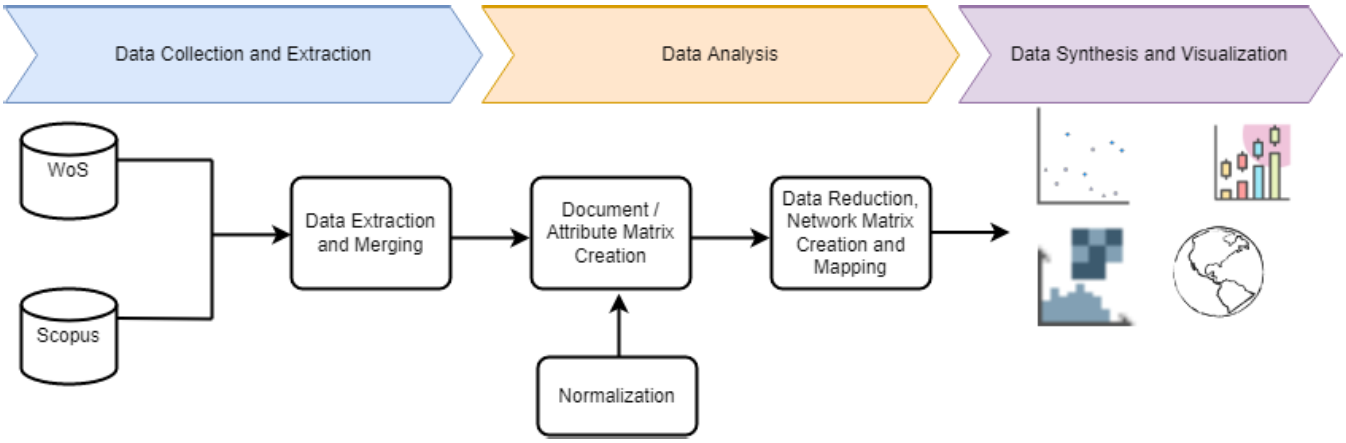

FIGURE 1. The bibliometric scientific mapping process [17].

TABLE 1. Selection criteria of the publications.

Criteria	Value
Source	WoS & Scopus
Search keywords (WoS)	(“Comput* education”) AND (“Retention” OR “Attrition” OR “Student retention” OR “Drop-out prevention” OR "Student success” OR "Support strategies” OR “Student integration”) OR (TOPIC) (“Comput* science”) AND (“Retention” OR “Attrition” OR “Student retention” OR “Drop-out prevention” OR “Student success” OR “Support strategies” OR “Student integration”)) OR TITLE-ABS-KEY (“Comput* engineering”) AND (“Retention” OR “Attrition” OR “Student retention” OR “Drop-out prevention” OR "Student success” OR "Support strategies” OR "Student integration”)
Search keywords (Scopus)	(TITLE-ABS-KEY ((“Comput* education”) AND (“Retention” OR “Attrition” OR “Student retention” OR “Drop-out prevention” OR "Student success” OR "Support strategies” OR “Student integration”)) OR TITLE-ABS-KEY ((“Comput* science”) AND (“Retention” OR “Attrition” OR “Student retention” OR “Drop-out prevention” OR “Student success” OR “Support strategies” OR “Student integration”)) OR TITLE-ABS-KEY ((“Comput* engineering”) AND (“Retention” OR “Attrition” OR “Student retention” OR “Drop-out prevention” OR "Student success” OR "Support strategies” OR "Student integration”)))
Search start date	January 1, 2000
Search end date	December 29, 2022
Document type	Multiple sources
Language	English
Number of articles	1310

is defined as the size of the intersection between two sets divided by the size of the union of the two sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where:

$A = x_1, x_2, x_3, \ldots, x_n$ where $x_i \in U$ and $i = 1, 2, \ldots, n$
$B = y_1, y_2, y_3, \ldots, y_m$ where $y_j \in U$ and $j = 1, 2, \ldots, m$

For example, Jaccard’s coefficient can be used to measure the similarity between two sets of articles based on the keywords they contain. If set A contains articles that include the keywords ‘‘diversity’’ and ‘‘attrition’’, and set B contains articles that include the keywords ‘‘inclusion’’ and ‘‘persistence’’, the Jaccard’s coefficient between A and B would be:

$$J(A, B) = \frac{|\text{‘‘diversity’’, ‘‘attrition’’} \cap \text{‘‘inclusion’’, ‘‘persistence’’}|}{|\text{‘‘diversity’’, ‘‘attrition’’} \cup \text{‘‘inclusion’’, ‘‘persistence’’}|} = \frac{0}{4} = 0 \tag{2}$$

Another similarity measure is the Salton’s cosine that takes into account the frequency of occurrence of each item in the two sets. It is a similarity measure between two non-zero vectors in an inner product space:

$$S(A, B) == \frac{\sum_{i=1}^{n}(a_i \cdot b_i)}{\sqrt{\sum_{i=1}^{n}(a_i^2)} \cdot \sqrt{\sum_{i=1}^{n}(b_i^2)}} \tag{3}$$

where:

a_i : represents the frequency of keyword i in set A.
b_i : represents the frequency of keyword i in set B.

For example, Salton’s cosine can be used to measure the similarity between two sets of articles based on the frequency of occurrence of specific keywords. If set A contains articles that mention the keyword ‘‘diversity’’ 10 times and the keyword ‘‘attrition’’ 5 times, and set B contains articles that mention ‘‘diversity’’ 8 times and ‘‘inclusion’’ 6 times, the Salton’s cosine between A and B would be:

Set A: $A = a_1, a_2$, where a_1 represents the frequency of ‘‘diversity’’ in set A, and a_2 represents the frequency of ‘‘attrition’’ in set A.

TABLE 2. Main Information about the attrition studies dataset.

Description	Results
Basic Information	
Period	2000:2022
Type of Source: Journal, conferences, books etc	416
Document	1310
Average years from publication	9.03
Average citations per documents	9.918
Average citations per year per document	0.9442
References	27350
Type of document	
Journal	232
Conference Paper	1029
Book Review	27
Note	1
Retracted	2
Review	19
Content of document	
Keyword plus	5054
Author's Keywords	2415
Authors	
Authors	3143
Author Appearances	4158
Authors of single-authored documents	156
Authors of multi-authored documents	2987
Authors Collaboration	
Single-authored documents	212
Document per Author	0.417
Authors per Document	2.4
Co-Authors per Documents	3.17
Collaboration Index	2.72

Set B: $B = b_1, b_2,$ where b_1 represents the frequency of "diversity" in set B, and b_2 represents the frequency of "attrition" in set B

Given values:

$$a_1 = 10 \text{ (frequency of "diversity" in set A)}$$
$$a_2 = 5 \text{ (frequency of "attrition" in set A)}$$
$$b_1 = 8 \text{ (frequency of "diversity" in set B)}$$
$$b_2 = 6 \text{ (frequency of "inclusion" in set B)}$$

Now we can substitute these values into the Salton's cosine similarity formula:

$$\text{Salton's cosine similarity} = \frac{\sum_{i=1}^{2}(a_i \cdot b_i)}{\sqrt{\sum_{i=1}^{2}(a_i^2)} \cdot \sqrt{\sum_{i=1}^{2}(b_i^2)}} \tag{4}$$

Substituting the given values:

$$= \frac{(10 \cdot 8) + (5 \cdot 0)}{\sqrt{(10^2) + (5^2)} \cdot \sqrt{(8^2) + (6^2)}} \tag{5}$$
$$= \frac{80}{\sqrt{100 + 25} \cdot \sqrt{64 + 36}} \tag{6}$$
$$= \frac{80}{\sqrt{125} \cdot \sqrt{100}} \tag{7}$$
$$= \frac{80}{11.18 \cdot 10} \tag{8}$$
$$= \frac{80}{111.8} \tag{9}$$
$$\approx 0.716 \tag{10}$$

Therefore, the Salton's cosine similarity between sets A and B is approximately 0.716. Both Jaccard's coefficient and Salton's cosine can be used to measure the similarity between sets of items, but they have different strengths and weaknesses depending on the nature of the data being analyzed. These measures help identify similarities and differences between different articles, which can be useful in identifying patterns and trends in the data.

C. DATA SYNTHESIS AND VISUALIZATION

Several diagrams were used to analyze and visualize the data, including histogram, maps, dendrogram, wordcloud, and treemap plots. These tools are commonly used in data analysis and visualization to help identify patterns and trends in the data and to communicate findings. One tool that was specifically used in this work is VOSviewer [36], which is a software tool for constructing and visualizing bibliometric networks. VOSviewer can be used to create keyword co-occurrence networks, which can help identify the main topics and trends to understand the trends of publications on attrition in computer science education.

III. RESULTS

This section presents the results of the analysis of data extracted from WoS and Scopus. The results include information on the number of documents, author information per document, article time-span, and commonly used keywords in attrition studies in the field of computer science.

A. TEMPORAL VIEW OF PUBLICATIONS

According to Section II-A, a total of 1310 articles focusing on students' attrition in computer science education, published between 2000 and 2022, were considered. Figure 2 demonstrates that publications steadily increased from 2018 with an annual growth rate of 10.5%, but then the rate of growth flattened in 2020. Notably, 114 articles were published in 2020, followed by 113 in 2019 and 111 in 2021. There was a slight drop from 2021 to 2022, and we are unsure whether this was due to the impact of COVID-19.

B. IMPACT ANALYSIS

This section presents an impact analysis that measures the top publication venues for both journals and conferences. Additionally, we present the top twenty published institutions and the prominent countries where these institutions are located. The following sections provide a summary of the impact analysis.

1) TOP 10 MOST RELEVANT JOURNALS AND CONFERENCES

The top 10 popular publication venues are presented in Table 3 and Table 4. From Table 3, the Elsevier Computers and Education journal appeared to be the top-ranked journal during the selected period considered in this study. This is closely followed by the ACM Transactions on Computing Education, the IEEE Access, and the Education and Information Technologies. For conference proceedings, the

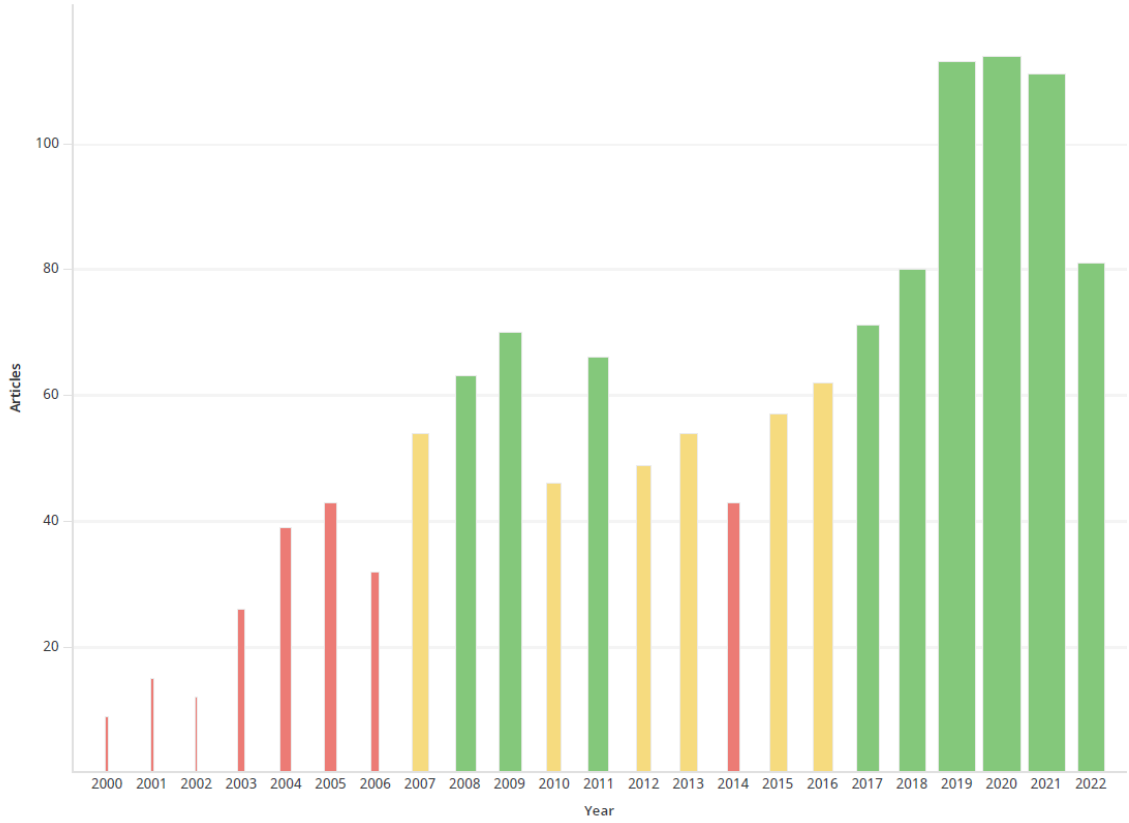

FIGURE 2. Time-plan of articles published between 2000 to 2022.

TABLE 3. Top ten journals.

Rank	Source
1	Elsevier Computers and Education
2	ACM Transactions on Computing Education
3	IEEE Access
4	IEEE Transactions on Education
5	Taylor and Francis Computer Science Education
6	Journal of Computer Assisted Learning
7	Education Technology and Society
9	IEEE Transactions on Learning Technologies
10	Education and Information Technologies

top ranked was the IEEE Frontiers in Education Conference (FIE), followed by the American Society for Engineering Education (ASEE) Annual Conference, and the ACM Symposium on Computer Science Education (SIGCSE). We observe that majority of the sources are devoted to computing and engineering education studies.

2) TOP TWENTY MOST PUBLISHED INSTITUTIONS

In relation to the top twenty most published institutions as shown in Table 5, we have observed that Arizona State University has the highest number of published articles with 44, followed by the University of California and Purdue University with 32 and 31 published articles respectively. Florida International University, Lamar University, and the University of Maryland are the last three institutions in the top twenty list. US institutions dominated the attrition studies list of universities, followed by Canada. In general, the list indicates that academic institutions in the US have a strong interest in this topic. One caveat to this data is that state universities such as the University of California have several campuses, each operating effectively as its own university. However, we were unable to separate by campus based on the data returned by our search. In contrast to the unified campuses of the University of California system, the University of Nebraska's campuses are entirely distinct within our analysis. One possible explanation for this discrepancy could be the lack of completeness in the data extraction process. Specifically, it is conceivable that the process did not retrieve all necessary metadata, such as unique campus identifiers or specific campus-related keywords, which are vital for accurately distinguishing between the different campuses of the University of California system. Moreover, inconsistencies in the naming conventions or variations used for the University of California campuses in the retrieved data might have caused confusion. For instance, abbreviations, acronyms, or local colloquial names used in place of the official campus names may not have been recognized by the analysis algorithm. Such irregularities in data extraction and processing might have led to an inaccurate aggregation of data, which, in turn, could have artificially inflated the ranking of certain universities in the list.

TABLE 4. Top ten conferences.

Rank	Source
1	IEEE Frontiers in Education Conference (FIE)
2	American Society for Engineering Education (ASEE) Conference
3	ACM Symposium on Computer Science Education (SIGCSE)
4	ACM Conference on Innovation and Technology in Computer Science Education (ITICSE)
5	IEEE GLOBAL Engineering Education Conference (EDUCON)
6	ACM Conference on International Computing Education Research (ICER)
7	IEEE Integrated STEM Education Conference
8	International Conference on Education and New Learning Technologies (EDULEARN)
9	International Conference on Computational Science and Computational Intelligence (CSCI)
10	International Technology Education and Development Conference

TABLE 5. Top twenty most published institutions.

Rank	Affiliations	Country	Articles
1	Arizona State University	USA	44
2	University of California	USA	32
3	Purdue University	USA	31
4	California State University	USA	29
5	University of Texas	USA	25
6	University of Virginia	USA	18
7	North Carolina State University	USA	16
8	University of Colorado	USA	16
9	University of Toronto	Canada	16
10	Simon Fraser University	Canada	15
11	University of Nevada	USA	18
12	Pennsylvania State University	USA	14
13	University of Florida	USA	14
14	University of Nebraska-Lincoln	USA	14
15	University of North Carolina at Charlotte	USA	14
16	Colorado State University	USA	13
17	Michigan State University	USA	13
18	Florida International University	USA	12
19	Lamar University	USA	12
20	University of Maryland	USA	12

TABLE 6. Top twenty most cited countries.

Rank	Country	Citations
1	USA	893
2	Canada	48
3	China	43
4	United Kingdom	32
5	Australia	28
6	Spain	19
7	Germany	16
8	India	15
10	Finland	14
11	South Africa	13
12	Turkey	12
13	Japan	11
14	Sweden	11
15	Ireland	10
16	Brazil	9
17	New Zealand	9
18	Saudi Arabia	9
19	South Korea	9
20	Italy	8

3) TOP 20 MOST CITED COUNTRIES

Table 6 shows the top twenty most cited countries. As illustrated in the table, the USA is in the first position, demonstrating its dominance in CS attrition research with 893 citations, followed by Canada with 48 citations, and China with 43 citations. Interestingly, European nations dominated the top twenty list. Among the Oceania nations, Australia had the highest number of citations with 28, followed by New Zealand with 9 citations. Notably, South Africa topped the list of African nations that appeared on the list with 13 citations.

C. CONCEPTUAL STRUCTURE

Della Corte et al. [40] use a conceptual structure to measure the quality of themes and to understand the evolution of a topic over time. This section provides a summary of the conceptual structure.

1) TOP THEMES OVER TIME

A thematic map is a tool that describes the conceptual structure of a particular study [40]. In Figure 3, we present such a map, which provides researchers with knowledge about the thematic areas of the study. The map displays a comprehensive evolution of sub-topics used over the years, with two key measures: centrality (on the x-axis) and density (on the y-axis). Centrality measures the level of inter-cluster interactions between topics, indicating how connected a given topic is to others. This helps researchers understand the overall coherence of the study's themes. Density, on the other hand, shows the intra-cluster relationship among keywords in a given theme and how they are developed over time. This measure is useful for understanding the level of detail and depth in the study's exploration of each theme.

The thematic map is divided into four quadrants: niche themes, motor theme, emerging themes, and basic themes. Niche themes have mostly unimportant external ties and marginal importance in the field. For example, themes such as ''learning environment'' and ''software engineering'' are relatively low or marginally important. The motor theme represents the main theme, characterized by high centrality and mostly conceptually related to other themes. The only theme within this category is ''motivation''. The emerging or declining themes are mostly weakly developed with low density, containing terms such as ''computer programming,'' etc. The last quadrant mostly contains basic themes, which are necessary for the field of CS attrition. Examples in this category are ''computer science education,'' ''active learning,'' ''retention,'' and ''computer science''.

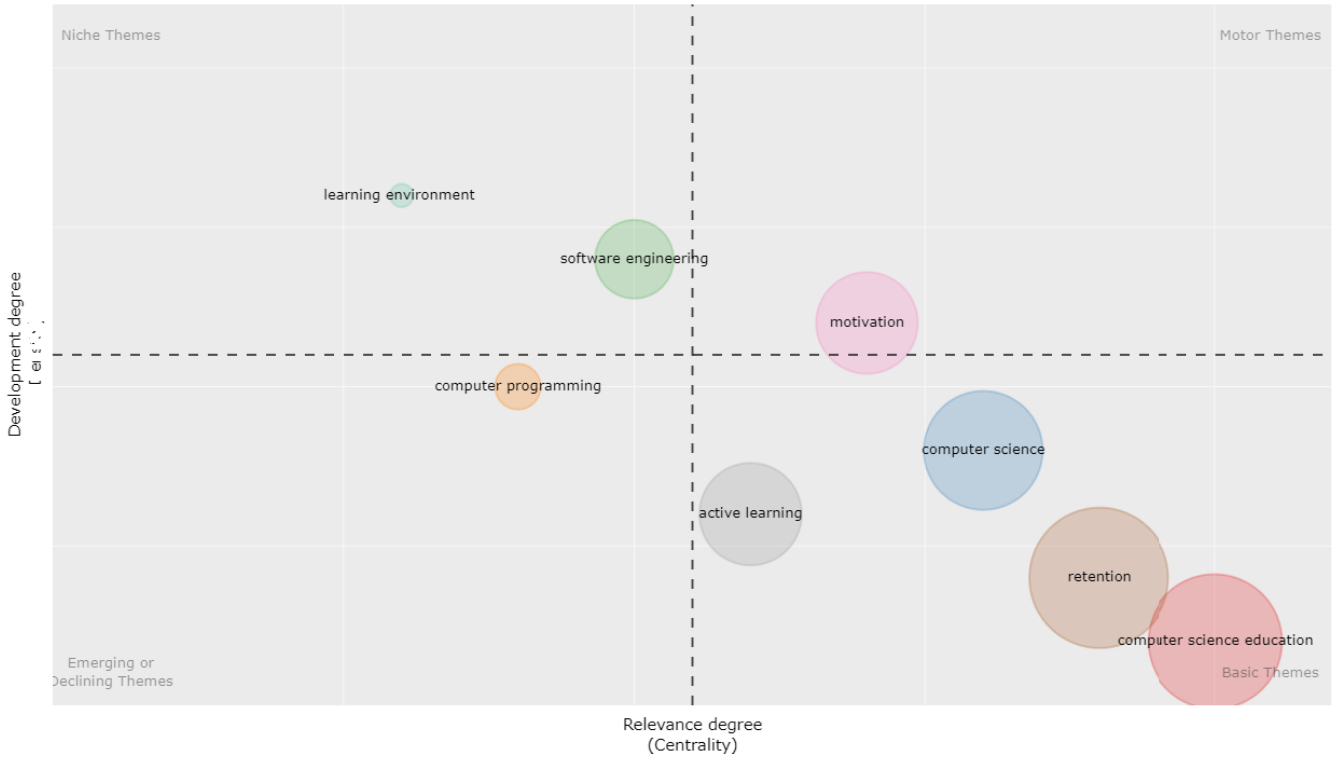

FIGURE 3. The trends of topics over time.

2) COMMON CO-OCCURRENCE NETWORK

To investigate the interconnection of commonly used topics, we use a co-occurrence network as shown in Figure 4. The co-occurrence keyword network can be constructed using the co-occurrence frequency of keywords in a set of documents. Once the co-occurrence frequency matrix has been constructed, a graph can be created where each keyword is represented by a node and the co-occurrence frequency between keywords is represented by the edges connecting the nodes. The strength of the relationship between two keywords can be represented by the thickness or weight of the edge. Several tools have utilized network analysis measures, such as degree centrality, betweenness centrality, and eigenvector centrality to analyze network structure and identify the most important or central keywords in the network.

Degree centrality is a measure that counts the number of direct connections a node has in a network [41], [42]. Mathematically, degree centrality of node v can be defined as:

$$C_D(v) = \frac{d_v}{n-1} \tag{11}$$

Here, $C_D(v)$ represents the degree centrality of node v, d_v represents the degree of node v, and n represents the total number of nodes in the network.

Betweenness centrality is a measure that quantifies the number of shortest paths between all pairs of nodes in a network that pass through a given node [43], [44]. Mathematically, betweenness centrality of node v can be defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{12}$$

Here, V is the set of nodes, $\sigma(s,t)$ is the total number of shortest paths from node s to node t, and $\sigma(s,t|v)$ is the number of those paths that pass through node v. The betweenness centrality of a node v is the sum of the fraction of all pairs of nodes that node v lies on the shortest path between.

Eigenvector centrality is a measure that takes into account the centrality of the nodes that are connected to a given node [45], [46]. Mathematically, eigenvector centrality of node i can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{j=1}^{n} a_{vj} x_j \tag{13}$$

Here, x_v represents the centrality score of node v, a_{vj} represents the edge weight between node v and node j, n is the total number of nodes in the network, and λ is a constant scalar value.

In the network, the size of the nodes represents the frequency of the topics, and the colored regions represent the commonly used topics. For instance, the common topics in the documents are "retention," "computer science education," "persistence," "recruitment," and "attrition".

`Biblioshiny` divides these topics into respective clusters, showing how these words are commonly used within a publication. For example, the red cluster shows the relationship between ''retention,'' ''mentoring,'' ''recruitment,'' ''pedagogy,'' ''attrition,'' ''higher education,'' ''persistence,'' ''engineering education,'' ''undergraduate research,'' ''curriculum,'' and ''broadening participation''.

D. KEYWORD VISUALIZATION

This section presents the common keywords used in CS attrition studies. To achieve this, we discuss the most frequently used keywords, strongly related keywords, and trend patterns.

1) TEMPORAL KEYWORD GROWTH

The graph in Figure 5 shows the annual distribution trend of keywords from 2000 to 2022. Identifying keyword trends can help researchers focus on newer or emerging areas in the field. Most keywords experienced a sharp increase throughout the years considered in our analysis. ''Retention'' remains the most popular keyword used by CS researchers, followed by ''computer science education'', ''computer science'', and ''computing education''. The choice of these keywords may be attributed to the countries in which the academic institutions are based. For example, US-based academic institutions prefer the ''computing education'' keyword, while European academic institutions use ''computer science education'' keywords. These keywords are likely to continue experiencing a significant increase in future trend analysis.

2) TOP-MOST KEYWORD FREQUENCY

To gain a better understanding of the most frequently used keywords, we have employed a word-cloud which displays the frequency of keywords used in a collection of publications. In a word cloud, the larger the keyword, the higher the frequency in the document. Figure 6 shows the commonly occurring keywords from our analysis, including ''computer science education'', ''retention'', ''computer science'', ''CS1'', and ''gender''. Smaller-sized keywords like ''self-efficacy'', ''e-learning'', ''learning analytics'', ''CS2'', ''data mining'', and ''inclusion'' are also present in the word-cloud, though they occur less frequently. These keywords are expected to continue to be prominent in the field and pave the way for future research directions.

3) TOP-MOST CO-OCCURRENCE KEYWORD NETWORK VISUALIZATION

Using the Vosviewer software, as depicted in Figure 7, we describe the top keywords used by authors. The closer two keywords are positioned to each other, the stronger their relatedness is, and the thickness of edges represents the strength of co-occurrence links between them. As presented in Figure 7, we selected the closely tied keywords or keywords with the highest total link strength out of 175 keywords. This indicates that these keywords are popular across authors. ''Retention'' was the most frequently occurring keyword, appearing 180 times with a total link strength of 365. ''Computer science education'' came next with 159 occurrences and 96 total link strength, followed by ''computer science'' with 100 occurrences and 57 total link strength.

4) TOP-MOST RELATED KEYWORDS

We used a tree-map to show the relatedness of keywords and to identify the most frequently used keywords. The percentage in the tree-map indicates the relevance of each keyword. As depicted in Figure 8, the tree-map shows that ''retention'' and ''computer science education'' are commonly used in research on CS student attrition. This indicates that these keywords are prominently used in the field. Other keywords of considerable interest that fall within this category are ''CS1,'' ''computer science,'' and ''gender.'' For keywords that are less related but still have significant relevance and are emerging, ''cs0'' and ''race'' are notable examples.

IV. DISCUSSION

In this section, we discuss the impact of our bibliometric analysis on our understanding of the landscape of retention studies in computing. From our observation, publications rose steadily from 2013 at 2.98% growth, with 2018 accounting for 226 articles but decreased suddenly to 105 publications between 2020 to 2021. We believe the COVID-19 pandemic might have impacted the number of articles produced (e.g. due to canceled conferences and decreased productivity of researchers). We do not believe that this decline reflects a decline in CS attrition. On the contrary, a recent study conducted by Mooney and Becker [47] mentioned that attrition rates increased as the COVID-19 pandemic continued. Similarly, Albarakati et al. [48] (2021) revealed that underrepresented minorities might even suffer attrition more significantly as a result of the pandemic.

In terms of where the articles have been published, our study reveals the importance of this topic across a range of communities. The top venues for this research include a set of journals and conferences that include both those that specifically focus on computing education (e.g. Elsevier Computers and Education, ACM Transactions on Computing Education) and more general education-focused venues (e.g. IEEE Transactions on Education, IEEE Frontiers in Education). This range of publication venues indicates that the problem of CS attrition is not just of narrow interest to those directly involved in the field of computing, but a problem that is important to the broader education community. Moreover the venues where this work is published are of high quality and impact. For example, the Computers and Education journal appeared to be the top-most published source with 24 publications by the year 2022. The Computers and Education journal is one of the leading journal in educational technology with a long publication history. According to Clarivate Analytics, a Web of Science Group, publishing annual report on journal citation, for 2022, Computers & Education journal was ranked second best with impact factor (IF) 11.25 which is next after the Review of Educational Research journal topmost with IF 13.55. Other journals and

FIGURE 4. Co-occurrence network of topics.

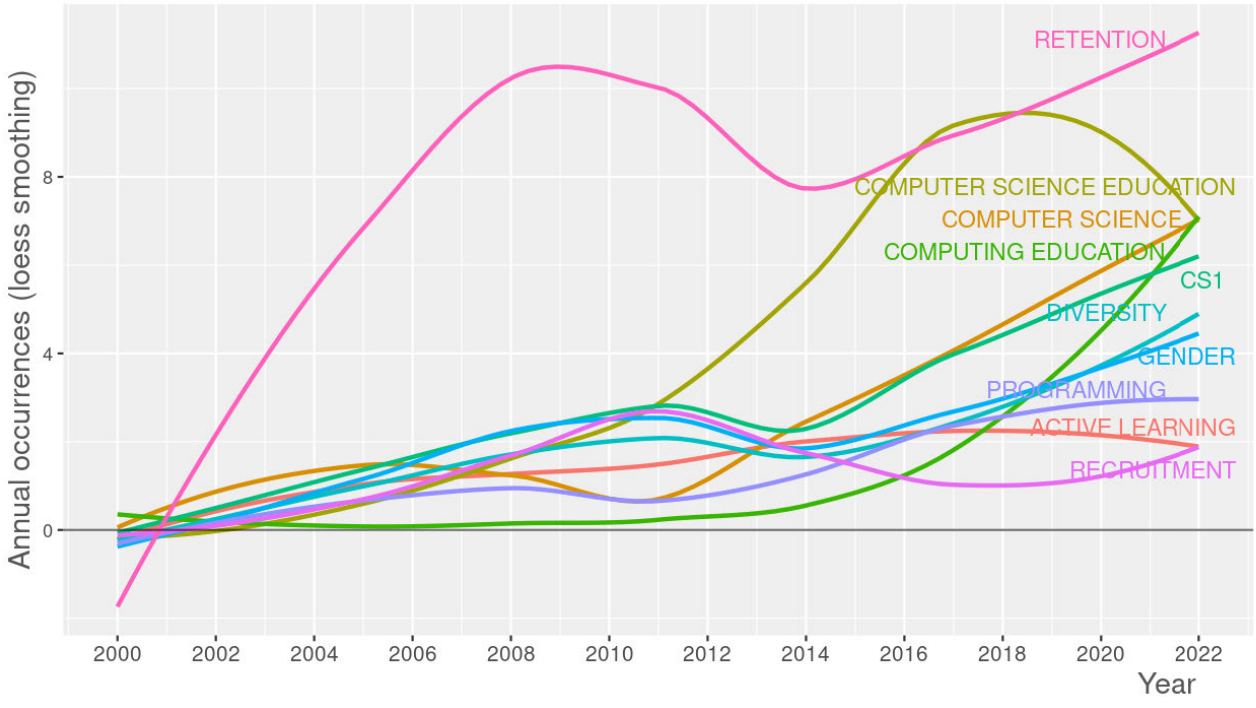

FIGURE 5. Temporal keyword growth from 2000 to 2022.

conferences on the list are also highly ranked. There may be other studies that are sent to lower-impact venues that do not appear in Scopus or the Web of Science which our study did not observe, but the presence of this work in high-impact

FIGURE 6. Top-most frequently occurring keywords.

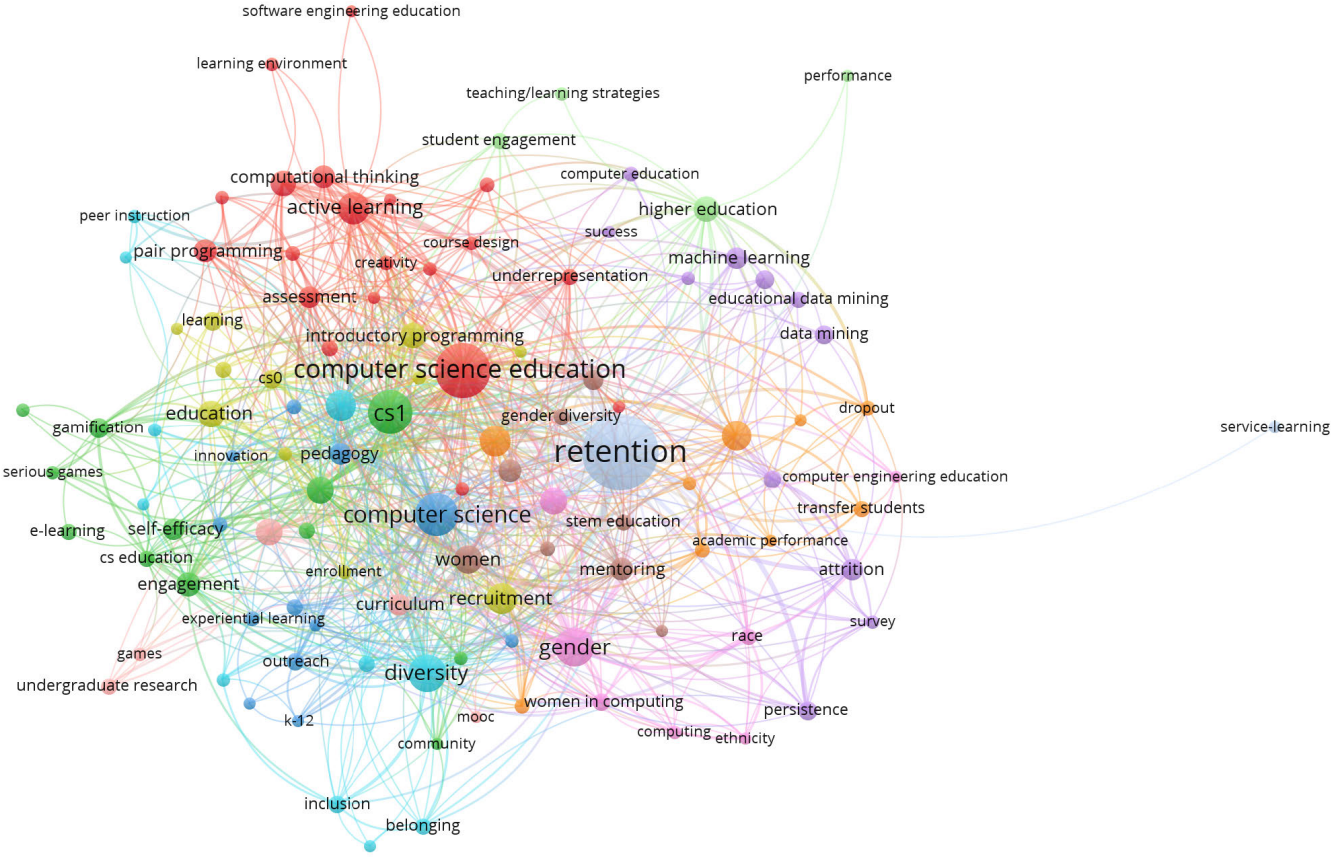

FIGURE 7. Top-most keyword interactions.

venues signals the growing importance of this topic in the field.

It is interesting to note that the United States dominates the publication count in this area. This is likely at least in part due to the size of the United States and the large number of US research universities. This trend does not imply that attrition is a bigger problem in the US compared to other countries, though our study cannot rule that out either.

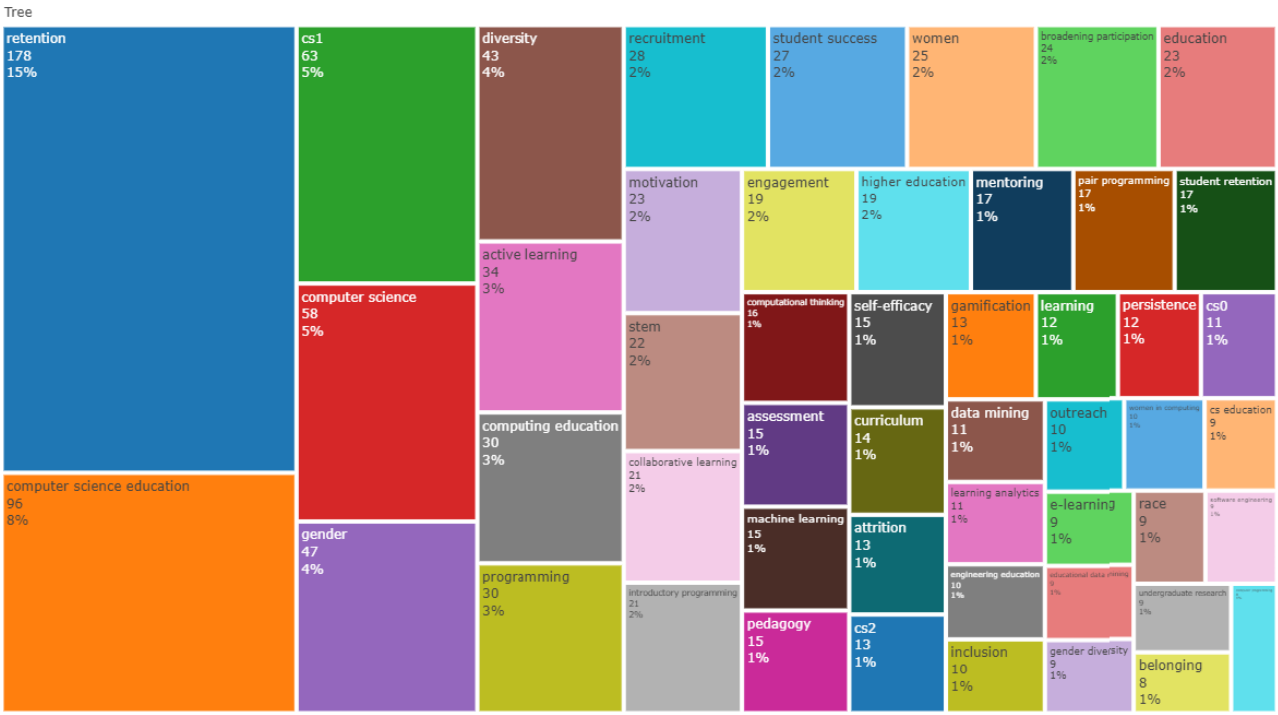

FIGURE 8. Top-most related keywords.

According to the conceptual structure analysis results, the four quadrants showing the trends of topic over time were considered. These quadrants consisted of niche, motor, basic, and emerging or declining themes. This study revealed that ''computer programming'' remain the only emerging theme in attrition studies focusing on computer science. While other themes such as ''active learning'' and ''computer science education'' are relevant and basic topics in attrition studies, ''computer programming'' seems to be the major focus connected to students'' attrition. This finding buttressed the overarching and known knowledge about how programming education is difficult for students to excel in and need more time and resources to thrive [8]. Therefore, students may dropout of pursuing a computing degree because they could not comprehend and develop computer programming skills due to its difficulty. The implication of this finding is that more interventions tailored towards facilitating students' programming education may be needed to improve retention in computer science as alluded by previous studies [1], [49].

The analysis of trending topics as revealed by the temporal keyword growth from 2000 to 2022 indicates that authors have focused their attention on student ''retention'' in the last two decades. Apparently, computer science researchers, for example, Cohoon [50], Cuny and Aspray [51], emphasised how to improve retention of female in computer science degree. Aside from computer science, other fields are showcasing research on student retention by investigating several characteristics with the aim to develop educational environments to foster students success [52], [53]. Thus, students' retention remain a hot research topic since it is relevant and critical to measuring educational success and whether academic strategies are meeting the learning needs of a given society [54]. Additionally, this study revealed that scholars researching students' attrition in computer science education are widening their scope by investigating computing education in general, including additional discipline-specific areas such as computer engineering, information systems, and software engineering [11].

V. CONCLUSION

This study provides insights into research on student attrition. It took a unique approach to qualitatively analyze relevant articles on student attrition through the lenses of bibliometric review study, with a focus on the context of computer science education. Our study serves as a guide to young and emerging researchers aiming to shape their research prospects on attrition in computer science education. For example, our analysis unveiled useful information about publication trends, venues, active institutions and countries, temporal keyword dynamics, and popular keywords used in this domain. With the findings of this study, young scholars in this field would have an overview of where to publish their relevant research, and how to position their research to focus on hot and trending topics. In addition, this study is relevant to scholars in this community as it visualizes the scientific progression witnessed in the field within the last two decades. For educators,

administrators, and other stakeholders, this study provides relevant information that can guide strategic planning and preparation for the future in order to address more issues related to student retention as a measure of academic success in the field.

This study is not without limitations. As is common with review studies, some of the actions taken to concretize the research procedure in accordance with the methodology adopted in this study may expose the study to certain limitations. For example, our bibliometric analysis excluded non-English articles. We acknowledge that this exclusion of articles written in languages other than English might lead to incomplete coverage of the study. However, due to the global recognition of English, this influence on our results might be insignificant. Another limitation is the lack of exhaustiveness of the data collected for our analysis, which the authors admit. Nevertheless, the extensive keywords used in conducting the search strategy for relevant data from two popular databases (Scopus and Web of Science) validate the measures taken to mitigate this limitation. In addition to the aforementioned limitations, it is important to note another significant constraint on our study. Both the University of Texas and the University of California are extensive public university systems with numerous campuses, departments, programs, and administrative bodies. However, due to their vastness and complexity, Scopus and Web of Science often conflate them into a single entity, leading to inaccuracies in research metrics and evaluations. As a result, the data we used to analyze the research productivity and impact of these universities may not fully capture the breadth and diversity of their contributions in our analysis. This limitation should be taken into consideration when interpreting our findings and drawing conclusions about the research performance of these institutions.

The study concludes by emphasizing the scientific progress made in the study of students' retention in computer science education as showcased by relevant research. Nevertheless, it draws the attention of stakeholders to the need for developing more strategies to create a niche in this domain, as current studies are primarily focused on computer programming while other areas of computing have received little attention.

REFERENCES

[1] C. Stephenson, A. D. Miller, C. Alvarado, L. Barker, V. Barr, T. Camp, C. Frieze, C. Lewis, E. C. Mindell, and L. Limbird, *Retention in Computer Science Undergraduate Programs in the U.S.: Data Challenges and Promising Interventions*. New York, NY, USA: ACM, 2018.

[2] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interact. Learn. Environ.*, vol. 30, no. 8, pp. 1414–1433, 2020.

[3] D. Bañeres, M. E. Rodríguez, A. E. Guerrero-Roldán, and A. Karadeniz, "An early warning system to detect at-risk students in online higher education," *Appl. Sci.*, vol. 10, no. 13, p. 4427, Jun. 2020.

[4] C. Isidro, R. M. Carro, and A. Ortigosa, "Dropout detection in MOOCs: An exploratory analysis," in *Proc. Int. Symp. Comput. Educ. (SIIE)*, Sep. 2018, pp. 1–6.

[5] V. Tinto, "Taking student success seriously: Rethinking the first year of college," *NACADA J.*, vol. 19, no. 2, pp. 5–9, 2005.

[6] J. B. G. Tilak and A. G. Kumar, "Policy changes in global higher education: What lessons do we learn from the COVID-19 pandemic?" *Higher Educ. Policy*, vol. 35, no. 3, pp. 610–628, Sep. 2022.

[7] M. Quadri and D. N. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global J. Comput. Sci. Technol.*, vol. 10, no. 2, pp. 1–8, 2010.

[8] P. Kinnunen and L. Malmi, "Why students drop out CS1 course?" in *Proc. 2nd Int. workshop Comput. Educ. Res.*, Sep. 2006, pp. 97–108.

[9] A. Tafliovich, J. Campbell, and A. Petersen, "A student perspective on prior experience in CS1," in *Proc. 44th ACM Tech. Symp. Comput. Sci. Educ.*, Mar. 2013, pp. 239–244.

[10] M. Xenos, C. Pierrakeas, and P. Pintelas, "A survey on student dropout rates and dropout causes concerning the students in the course of informatics of the Hellenic Open University," *Comput. Educ.*, vol. 39, no. 4, pp. 361–377, Dec. 2002.

[11] R. T. Palmer, D. C. Maramba, and T. E. Dancy, "A qualitative investigation of factors promoting the retention and persistence of students of color in STEM," *J. Negro Educ.*, vol. 80, no. 4, pp. 491–504, 2011.

[12] A. Reed, "Exploring the perceptions of current computer science students and student affairs professionals on the factors influencing student retention and attrition," Ph.D. dissertation, Texas Tech Univ., Lubbock, TX, USA, 2016.

[13] T. Beaubouef and J. Mason, "Why the high attrition rate for computer science students: Some thoughts and observations," *ACM SIGCSE Bull.*, vol. 37, no. 2, pp. 103–106, Jun. 2005.

[14] R. M. Powell, "Improving the persistence of first-year undergraduate women in computer science," *ACM SIGCSE Bull.*, vol. 40, no. 1, pp. 518–522, Feb. 2008.

[15] S. Sharmin, "Creativity in CS1: A literature review," *ACM Trans. Comput. Educ.*, vol. 22, no. 2, pp. 1–26, Jun. 2022.

[16] J. M. Keller, "Development and use of the ARCS model of instructional design," *J. Instructional Develop.*, vol. 10, no. 3, pp. 2–10, Sep. 1987.

[17] M. Aria and C. Cuccurullo, "Bibliometrix: An R-tool for comprehensive science mapping analysis," *J. Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017.

[18] F. J. Agbo, I. T. Sanusi, S. S. Oyelere, and J. Suhonen, "Application of virtual reality in computer science education: A systemic review based on bibliometric and content analysis methods," *Educ. Sci.*, vol. 11, no. 3, p. 142, Mar. 2021.

[19] F. J. Agbo, S. S. Oyelere, J. Suhonen, and M. Tukiainen, "Scientific production and thematic breakthroughs in smart learning environments: A bibliometric analysis," *Smart Learn. Environ.*, vol. 8, no. 1, pp. 1–25, Dec. 2021.

[20] B. Godin, "On the origins of bibliometrics," *Scientometrics*, vol. 68, no. 1, pp. 109–133, Jul. 2006.

[21] W. W. Hood and C. S. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics*, vol. 52, no. 2, pp. 291–314, 2001.

[22] F. J. Martínez-López, J. M. Merigó, L. Valenzuela-Fernández, and C. Nicolás, "Fifty years of the European journal of marketing: A bibliometric analysis," *Eur. J. Marketing*, vol. 52, nos. 1–2, pp. 439–468, Feb. 2018.

[23] J. Li, F. Goerlandt, and G. Reniers, "An overview of scientometric mapping for the safety science community: Methods, tools, and framework," *Saf. Sci.*, vol. 134, Feb. 2021, Art. no. 105093.

[24] A. Darko, A. P. C. Chan, M. A. Adabre, D. J. Edwards, M. R. Hosseini, and E. E. Ameyaw, "Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities," *Autom. Construct.*, vol. 112, Apr. 2020, Art. no. 103081.

[25] O. Ellegaard and J. A. Wallin, "The bibliometric analysis of scholarly production: How great is the impact?" *Scientometrics*, vol. 105, no. 3, pp. 1809–1831, Dec. 2015.

[26] A. Bakri and P. Willett, "Computer science research in Malaysia: A bibliometric analysis," in *Aslib Proceedings*. Bingley, U.K.: Emerald Group Publishing Limited, 2011.

[27] W. M. Sweileh, "Global research activity on E-learning in health sciences education: A bibliometric analysis," *Med. Sci. Educator*, vol. 31, no. 2, pp. 765–775, Apr. 2021.

[28] C. Madden, R. O'Malley, P. O'Connor, E. O'Dowd, D. Byrne, and S. Lydon, "Gender in authorship and editorship in medical education journals: A bibliometric review," *Med. Educ.*, vol. 55, no. 6, pp. 678–688, Jun. 2021.

[29] C.-C. Chao, J.-M. Yang, and W.-Y. Jen, "Determining technology trends and forecasts of RFID by a historical review and bibliometric analysis from 1991 to 2005," *Technovation*, vol. 27, no. 5, pp. 268–279, May 2007.
[30] V. Servantie, M. Cabrol, G. Guieu, and J.-P. Boissin, "Is international entrepreneurship a field? A bibliometric analysis of the literature (1989–2015)," *J. Int. Entrepreneurship*, vol. 14, no. 2, pp. 168–212, Jun. 2016.
[31] A. M. H. Abbas, K. I. Ghauth, and C. Ting, "User experience design using machine learning: A systematic review," *IEEE Access*, vol. 10, pp. 51501–51514, 2022.
[32] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "SciMAT: A new science mapping analysis software tool," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 8, pp. 1609–1630, Aug. 2012.
[33] N. J. Van Eck and L. Waltman, "Vosviewer manual," *Leiden, Univeristeit Leiden*, vol. 1, no. 1, pp. 1–53, 2013.
[34] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.
[35] Y. Hou and Z. Yu, "A bibliometric analysis of synchronous computer-mediated communication in language learning using VOSviewer and CitNetExplorer," *Educ. Sci.*, vol. 13, no. 2, p. 125, Jan. 2023.
[36] N. J. van Eck and L. Waltman, "Citation-based clustering of publications using CitNetExplorer and VOSviewer," *Scientometrics*, vol. 111, no. 2, pp. 1053–1070, May 2017.
[37] Y. Zhang, L. Shang, L. Huang, A. L. Porter, G. Zhang, J. Lu, and D. Zhu, "A hybrid similarity measure method for patent portfolio analysis," *J. Informetrics*, vol. 10, no. 4, pp. 1108–1130, Nov. 2016.
[38] C. Sternitzke and I. Bergmann, "Similarity measures for document mapping: A comparative study on the level of an individual scientist," *Scientometrics*, vol. 78, no. 1, pp. 113–130, Jan. 2009.
[39] P. Glenisson, W. Glänzel, F. Janssens, and B. D. Moor, "Combining full text and bibliometric information in mapping scientific disciplines," *Inf. Process. Manage.*, vol. 41, no. 6, pp. 1548–1572, Dec. 2005.
[40] V. Della Corte, G. Del Gaudio, F. Sepe, and F. Sciarelli, "Sustainable tourism in the open innovation realm: A bibliometric analysis," *Sustainability*, vol. 11, no. 21, p. 6114, Nov. 2019.
[41] J. Zhang and Y. Luo, "Degree centrality, betweenness centrality, and closeness centrality in social network," in *Proc. 2nd Int. Conf. Modeling, Simulation Appl. Math. (MSAM)*, 2017, pp. 300–303.
[42] P. Bródka, K. Skibicki, P. Kazienko, and K. Musial, "A degree centrality in multi-layered social network," in *Proc. Int. Conf. Comput. Aspects Social Netw. (CASoN)*, Oct. 2011, pp. 237–242.
[43] M. Barthelemy, "Betweenness centrality in large complex networks," *Eur. Phys. J. B Condens. Matter*, vol. 38, no. 2, pp. 163–168, Mar. 2004.
[44] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.
[45] P. Bonacich, "Some unique properties of eigenvector centrality," *Social Netw.*, vol. 29, no. 4, pp. 555–564, Oct. 2007.
[46] B. Ruhnau, "Eigenvector-centrality—A node-centrality?" *Social Netw.*, vol. 22, no. 4, pp. 357–365, Oct. 2000.
[47] C. Mooney and B. A. Becker, "Investigating the impact of the COVID-19 pandemic on computing students' sense of belonging," *ACM Inroads*, vol. 12, no. 2, pp. 38–45, Jun. 2021.
[48] N. Albarakati, L. DiPippo, and V. Fay-Wolfe, "Rethinking CS0 to improve performance and retention," in *Proc. Australas. Comput. Educ. Conf.*, Feb. 2021, pp. 131–137.
[49] K. Pantic and J. Clarke-Midura, "Factors that influence retention of women in the computer science major: A systematic literature review," *J. Women Minorities Sci. Eng.*, vol. 25, no. 2, pp. 119–145, 2019.
[50] J. M. Cohoon, "Toward improving female retention in the computer science major," *Commun. ACM*, vol. 44, no. 5, pp. 108–114, May 2001.
[51] J. Cuny and W. Aspray, "Recruitment and retention of women graduate students in computer science and engineering: Results of a workshop organized by the computing research association," *ACM SIGCSE Bull.*, vol. 34, no. 2, pp. 168–174, Jun. 2002.
[52] L. E. Bernold, J. E. Spurlin, and C. M. Anson, "Understanding our students: A longitudinal-study of success and failure in engineering with implications for increased retention," *J. Eng. Educ.*, vol. 96, no. 3, pp. 263–274, Jul. 2007.
[53] M. Parker, "Placement, retention, and success: A longitudinal study of mathematics and retention," *J. Gen. Educ.*, vol. 54, no. 1, pp. 22–40, Jan. 2005.
[54] G. D. Caruth, "Student engagement, retention, and motivation: Assessing academic success in today's college students," *Participatory Educ. Res.*, vol. 5, no. 1, pp. 17–30, Dec. 2018.

GEORGE OBAIDO (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer science from the University of the Witwatersrand, Johannesburg, South Africa. He is currently a Post-Doctoral Scholar with the Center for Human-Compatible Artificial Intelligence (CHAI), Berkeley Institute for Data Science (BIDS), University of California at Berkeley, Berkeley, CA, USA. His research interest includes finding solutions to many societal problems using machine learning.

FRIDAY JOSEPH AGBO received the Ph.D. degree in computer science from the University of Eastern Finland. He was a University Lecturer with the School of Computing, University of Eastern Finland, Joensuu, Finland. He is currently an Assistant Professor of computer science at the School of Computing and Information Sciences, Willamette University. His research interests include broadening participation in computer science, designing and developing smart learning environments to foster novices' understanding, and gaining computational skills, including problem-solving, computational thinking, and programming education in general, using virtual reality technology, grounded in experiential learning theory, and game-based learning for a 21st-century learning experience.

CHRISTINE ALVARADO is currently the Associate Dean with the Division of Undergraduate Education, University of California at San Diego, where she is also a Teaching Professor and holds the Paul R. Kube Chair with the Department of Computer Science and Engineering (CSE). She is also with the Center for Inclusive Computing as the Program Manager for the Transfer Pathways Program. Her current research interest includes designing curriculum and programs to make computing and computing education more accessible and appealing, with an emphasis on increasing the number of women, Black, Latinx, Native American, and Pacific Islander students who study computing.

SOLOMON SUNDAY OYELERE received the B.Tech. degree (Hons.) in computer science from the Federal University of Technology Yola, Nigeria, the M.Sc. degree (Research) in computer and systems engineering from the Ilmenau University of Technology, Germany, and the Ph.D. degree in computer science from the University of Eastern Finland. He is currently an Associate Professor with the Luleå University of Technology, Sweden. His research interests include mobile and context-aware computing, smart learning environments, and pervasive and interactive systems. His current research interests include developing smart technology and games to support education and healthcare.