Contemporary Educational Technology, 2024, **16**(2), ep502
ISSN: 1309-517X (Online)

OPEN ACCESS

Review Article

Bibliometric insights into data mining in education research: A decade in review

Yessane Shrrie Nagendhra Rao [1]
0009-0002-9283-2922

Chwen Jen Chen [1*]
0000-0001-5175-7060

[1] Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Sarawak, MALAYSIA
[*] Corresponding author: cjchen@unimas.my

Citation: Rao, Y. S. N., & Chen, C. J. (2024). Bibliometric insights into data mining in education research: A decade in review. *Contemporary Educational Technology, 16*(2), ep502. https://doi.org/10.30935/cedtech/14333

ARTICLE INFO

Received: 17 Nov 2023
Accepted: 7 Feb 2024

ABSTRACT

This bibliometric study on data mining in education synonymous with big educational data utilizes VOSviewer and Harzing's Publish and Perish to analyze the metadata of 1,439 journal articles found in Scopus from 2010 to 2022. As bibliometric analyses in this field are lacking, this study aims to provide a comprehensive outlook on the current developments and impact of research in this field. This study employs descriptive and trends analysis, co-authorship analysis, co-citation analysis, co-occurrences of keywords, terms map analysis, and analysis of the impact and performance of publications. It also partially replicates a similar study conducted by Wang et al. (2022), who used the Web of Science (WoS) database. The study is reported in an article entitled 'Big data and data mining in education: A bibliometrics study from 2010 to 2022'. Results show that data mining in education is a growing research field. There is also a significant difference between the publications in Scopus and WoS. The study found several research areas and topics, such as student academic performance prediction, e-learning, machine learning, and innovative data mining techniques, to be the core basis for collaborating and continuing current research in this field. These results highlight the importance of continuing research on data mining in education, guiding future research in tackling educational challenges.

Keywords: educational data mining, big data, education, bibliometric analysis, Scopus

INTRODUCTION

Technology is no longer considered new and unconventional (Rodrigues et al., 2018). The fast-growing tendency of technology deliberately leads to more systematic and sophisticated ways of capturing users' data. These data are procured in large sets that cannot be interpreted thoroughly through conventional methods and are currently referred to as big data (Marín-Marín et al., 2019; Sin & Muthu, 2015). Nevertheless, various data mining techniques have been used to analyze and interpret the data collected as much as possible to influence practices, procedures, and decision-making in diverse fields, including the field of education (Baek & Doleck, 2022; Menon et al., 2017; Rodrigues et al., 2018).

Data Mining & Big Data in Education

According to International Educational Data Mining Society (2011), data mining in education or educational data mining (EDM) is defined as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to understand students better, and the settings, which they learn in." Data mining in education has been a growing research area among many researchers (Baek & Doleck, 2022; Romero & Ventura, 2010). Researchers are interested in using data mining techniques and algorithms on big data to analyze and derive relevant and useful educational insights (Romero & Ventura, 2010).

Romero and Ventura (2010) have identified several contexts in which EDM has and can be used to "convert the raw data coming from the educational systems into useful information" (p. 601). The contexts include the traditional classroom setting, where educators and learners are seated in a face-to-face learning environment, blended learning environments, e-learning environments, learning management systems (LMS), intelligent tutoring systems (ITS), massive open online course (MOOC), and adaptive educational hypermedia systems (Romero & Ventura, 2010, 2017). Apart from data mining, Romero and Ventura (2010) highlighted that web mining techniques are also used in the context of online systems by applying them to students' data, which is stored in the underlying databases and log files of the users.

According to Fischer et al. (2020), the emergence of the digital era has brought about the emergence of big data, where students' data in a traditional setting have now come to be digitized through the student information system) and also procured through the increased implementation of LMS. Masood and Mokmin (2017) have also highlighted the use of ITS with big data, which can provide a positive impact on the process of teaching and learning (T&L) with the help of the Internet of things (IoT) technologies. A variety of data mining techniques can be applied to the said data, which can generally be categorized into prediction methods, including inferential techniques that bring upon a paradigm shift in transforming the dynamic information obtained through these data into adaptive knowledge, structure discovery algorithms, with an emphasis on exploring the edifices of content and competencies in an educational setting, as well as the internal structures of learners' social networks; relationship mining, including correlation and sequential pattern mining; and visual analytics (Baker & Siemens, 2014; Fischer et al., 2020).

Bibliometric Analysis & Previous Related Studies

This study employed the bibliometrics analysis (BA) method to explore and examine research development over the past decade in the context of data mining in education. BA is a quantitative method in research used to analyze the properties of the metadata of academic publications, which are relevant to the scope of study of a researcher while also describing the development trends of the said research field (Ahmi, 2021; Wang et al., 2022). It is a popular research method in many fields. However, the publications on BA in data mining in education are lacking or are more or less combined with systematic literature reviews (SLR). With EDM being a multidisciplinary field, fragmented into different research areas, BA research method is useful for analyzing the common areas, where EDM studies are conducted (Baek & Doleck, 2022; Hermaliani et al., 2022). BA allows for the fundamental understanding of EDM (Baek & Doleck, 2022). On the other hand, with big data becoming a widely researched concept, BA can be used to acknowledge further the impact of big data in education through the scholarly articles and publications related to it (Marín-Marín et al., 2019).

Marín-Marín et al. (2019) conducted a bibliometric study encompassing four databases, which were Web of Science (WoS), Scopus, Education Research Information Center (ERIC), and PsycINFO, spanning from 2010 to 2018. Marín-Marín et al. (2019) found that there has been increased publication on big data in education in 2017, offering valuable insights into referent authors, countries, and high statistical value and mapping. Marín-Marín et al. (2019) also found that publications on big data in education started in 2010, underlining the trending and promising nature of big data research and its relevance in evaluating T&L processes. Another study on BA conducted by Baek and Doleck (2022) using the Scopus database from 2015 to 2019 shows that the primary goals of current EDM research are to create and enhance methods for data analysis using cutting-edge technologies and incorporate pedagogy and learning contexts (Baek & Doleck, 2022). Wang et al. (2022) discovered that a major contribution to the literature on EDM and big data in their BA were from the educational technology and computer education fields. They used the WoS database to retrieve the selected articles. The results of BA show rapid growth in research on data mining and big data over the past decade. (Wang et al., 2022).

BA method has also been used relatively more with SLR. Among 291 publications derived from Scopus with approximate keywords, bibliometric analysis, SLR, EDM, and big data in education, 288 of the publications utilized the review methodology. Only three publications independently adopted BA method as of December 22, 2022, which are the earlier described Baek and Doleck (2022), Marín-Marín et al. (2019), and Wang et al. (2022). One review, however, did include BA as a secondary method to gauge development trends. Hermaliani et al. (2022) conducted a systematic review and bibliometric study on how EDM has evolved in predicting student performances from 2015 to 2021 and which data mining techniques were used using the Scopus

database. Findings indicate machine learning algorithms' surge in student performance prediction, reflecting positive growth over the years (Hermaliani et al., 2022).

Research Aim & Questions

The studies conducted by Baek and Doleck (2022), Hermaliani et al. (2022), Marín-Marín et al. (2019), and Wang et al. (2022) all highlight that further bibliometric analyses need to be conducted in the field of data mining in education using different databases. The present study partially replicates the study conducted by Wang et al. (2022), who utilized the WoS database. This study used another large academic publication database, Scopus, to conduct similar yet detailed research regarding EDM and big data. It compared findings between the literature found in WoS and Scopus accordingly. This study generally aims to analyze the scientific productivity of data mining in the context of education between 2010 and 2022. The following questions were established to achieve the aim of this study:

1. Which countries, institutions, and authors represent the preponderance of academic publications in the field of data mining in education?
2. What are the collaborative zones of literature, namely co-authorship, co-citations, and co-occurrences of keywords identified?
3. What is the impact and performance of the articles and citations of a similar subject and context on data mining in education?
4. What are the similarities and differences between the findings of this study through the academic publications of Scopus compared to that of WoS as researched by Wang et al. (2022)?
5. What is the evolution of research topics over the years?

In this study, big data in education is apprehended as synonymous with data mining in education. Hence, the essence and the use of data mining in education will include big educational data or big data in education.

MATERIALS & METHOD

Research Method

Bibliometric analysis is the only method used in this study to analyze the retrieved academic publications from 2010 to 2022. The academic publications were obtained from the Scopus database developed by Elsevier. The Scopus search result analysis is used to summarize the descriptive statistics of the dataset, which will show the trend analysis of the academic publications and answer research question 1. The tools used to conduct the bibliometric analysis are VOSviewer and Harzing's Publish or Perish. The first tool was used to visualize the analyzed metadata of academic publications in terms of co-citation, co-authorship, co-occurrences, and terms map analysis, which answers research question 2 and question 5, while the latter was used to answer research question 3 in getting information from academic publications such as "total citations (TC), citation per paper (C/P), number of cited papers (NCP), citations per cited paper (C/CP), g-index, and h-index" (Ahmi, 2021, p. 81).

Research question 4 was answered in terms of the findings of research questions 1, 2, and 3. The findings will highlight the similarities and differences found between this study and that of Wang et al. (2022). These could be learned in terms of the descriptive and trend analysis components, such as seeing through the differences in the collected data under the unit of analysis of the dataset columns, the impact and performance components such as TC, C/P, NCP, C/CP, total publications (TP), citations per year (C/Y), h-index, and g-index and through the visualization and mapping of co-occurrences of keywords, co-citation, and co-authorship based on the authors, sources, and countries.

Data Source & Retrieval

As shown in the modified PRISMA framework in **Figure 1**, a total of 4,681 journal articles were retrieved from the Scopus Database using the search string '(TITLE-ABS-KEY ("educational data mining" OR "data mining in education" OR "educational big data") OR TITLE-ABS-KEY (education AND ("big data" OR "data mining"))) AND (PUBYEAR>2009) AND (PUBYEAR<2023) AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English"))'.

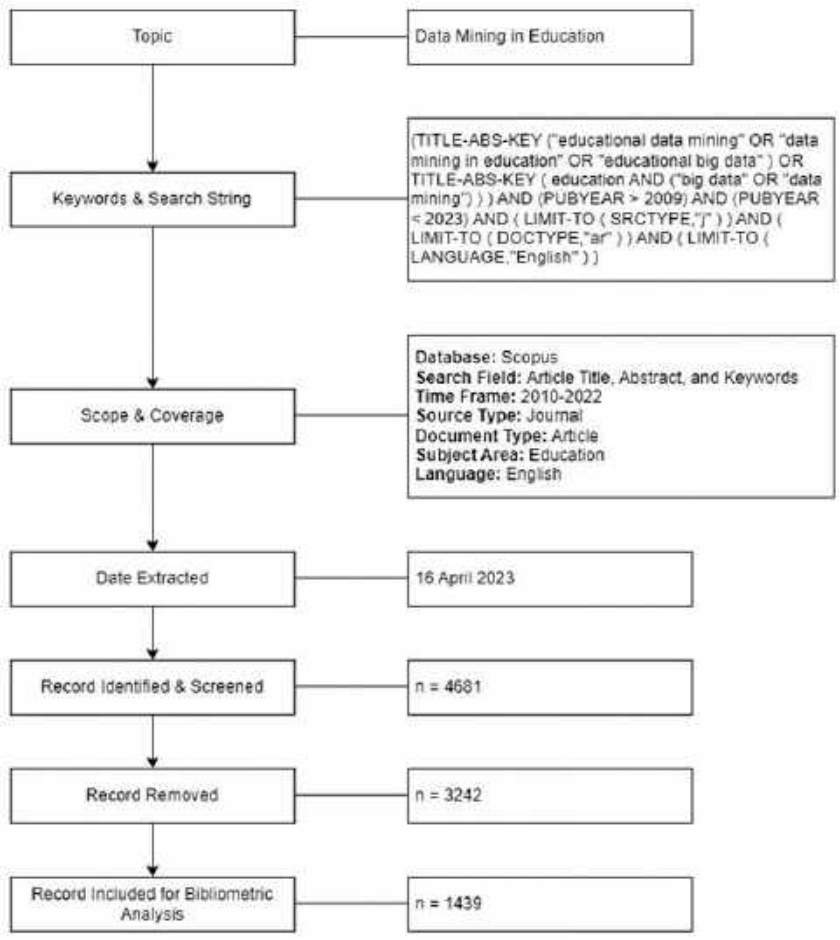

Figure 1. Modified PRISMA framework (Source: Authors, adopted from Ahmi, 2021)

The search string sets the retrieved data to be that of English journal articles within the period of 2010 and 2022, which are within the research scope of data mining in education and that of big data. However, several journal articles were connected through the same keywords that are unrelated or inconsistent with the subject of this BA study. The literature data obtained has to be pre-processed accordingly to avoid irrelevant results that may affect the findings' quality.

The method used to pre-process the data was by saving the search results into three different lists, namely results from 2010 to 2016, 2017 to 2020, and 2021 to 2022, and manually screening the articles by deleting the articles that were not necessary from the lists. This method of manual screening was proposed by Ahmi (2021). If the same authors publish any duplicate studies or different versions of studies, the most recent version will be selected. Hence, after a thorough manual screening of the initial number of articles, 1439 valid journal publications were obtained and used for the analysis.

Inclusion & exclusion criteria

The chosen literature papers only included journal articles on empirical studies or systematic reviews. Both published and in-press journal articles were included. However, this study excluded journal articles on bibliometric studies and studies not within the educational context.

RESULTS

Publications by Years

Figure 2 shows the total academic publications in data mining in education from 2010 to 2022. There was a rapid increase from 2015 to 2018 and a big leap in 2019. The publications continued to increase from 2019 to 2020 but had a minor fallback of 33 publications in 2021.

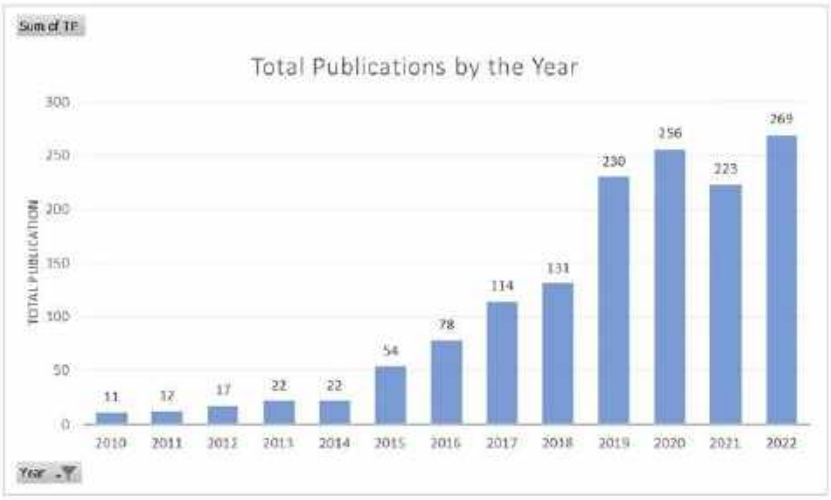

Figure 2. Distribution of publications from 2010 to 2022 (Source: Authors, using data that is obtained through the search string constructed by the authors in Scopus)

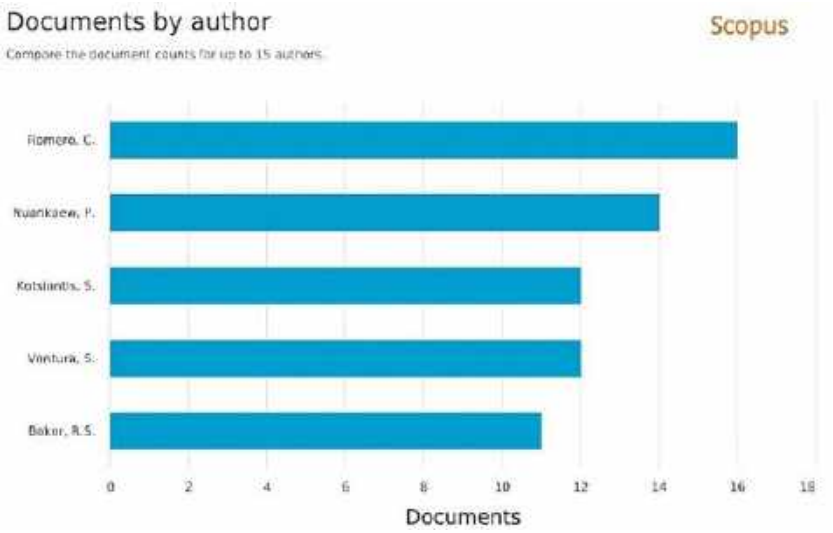

Figure 3. Top-5 authors in data mining in education (Source: Authors, chart generated through Scopus)

In 2022, as of the results obtained on April 16, 2023, the publications have once again increased, with it being the year with the most publications, 269 journal articles. This shows that despite the minor fallback in the year 2021, the research on EDM and big educational data is still a research area currently being focused on.

Representation of Authors, Institutions, & Countries in Data Mining in Education

Representation of authors

Figure 3 shows the total publications of the top-5 authors in data mining in education from 2010 to 2022. The author at the top-spot is Romero, C., who has been a large contributor to research publications focused on EDM. The author has a total of 16 research publications from 2011 to 2022. The authors, Nuankaew, P., Kotsiantis, S., and Ventura, S., are in the second, third, and fourth places, respectively. Nuankaew, P. has 14 documents, whereas Kotsiantis, S. and Ventura, S. have 12 publications each. Romero, C. and Ventura, S. are influential authors from the same country, Spain. The final author in the top-5 list is Baker, S., with 11 publications.

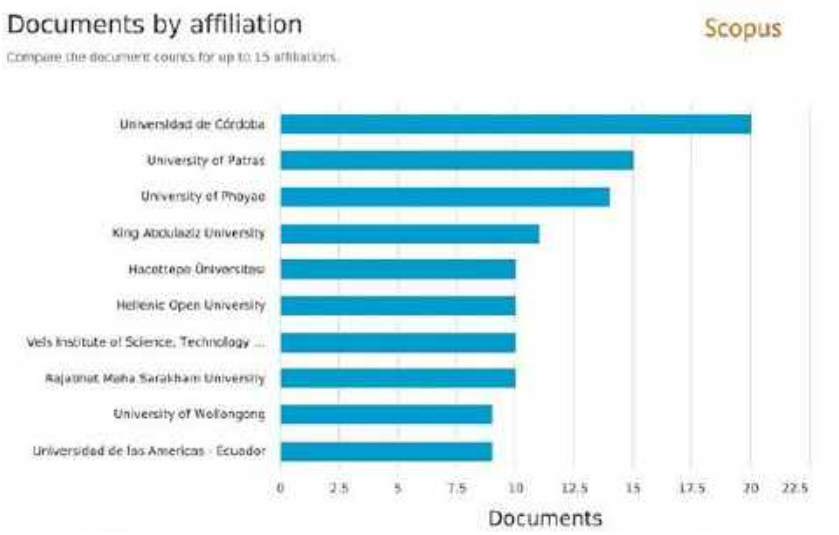

Figure 4. Top-10 institutions in data mining in education (Source: Authors, chart generated through Scopus)

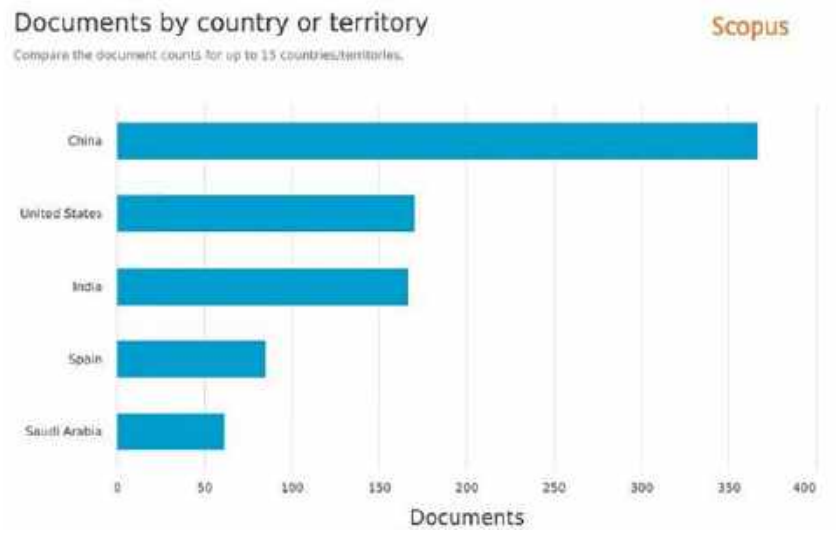

Figure 5. Top-5 countries in data mining in education (Source: Authors, chart generated through Scopus)

Representation of institutions

Figure 4 shows the top-10 institutions with the highest number of journal articles published in this field of study.

The University of Cordoba (Universidad de Córdoba) leads with 20 journal articles, followed by the University of Patras, the University of Phayao, and the King Abdulaziz University with 15, 14, and 11 publications. Hacettepe University (Hacettepe Üniversitesi), Hellenic Open University, Vels Institute of Science, Technology & Advanced Studies, and Rajabhat Maha Sarakham University have 10 publications, respectively. The University of Wollongong and the University of the Americas–Ecuador (Universidad de las Americas–Ecuador) follow closely with nine publications. Notably, influential authors in this field, Romero, C. and Ventura, S. from the University of Cordoba, have significantly contributed, often collaboratively, to these publications.

Representation of countries

According to **Figure 5**, the top-5 countries that have contributed to academic research in the field of data mining in education are China, The United States, India, Spain, and Saudi Arabia. With 366 papers, China has had the most publications, surpassing The United States by 196 papers. The articles published by authors from China are dated from 2010 to 2022 in Scopus. There is little difference between the number of

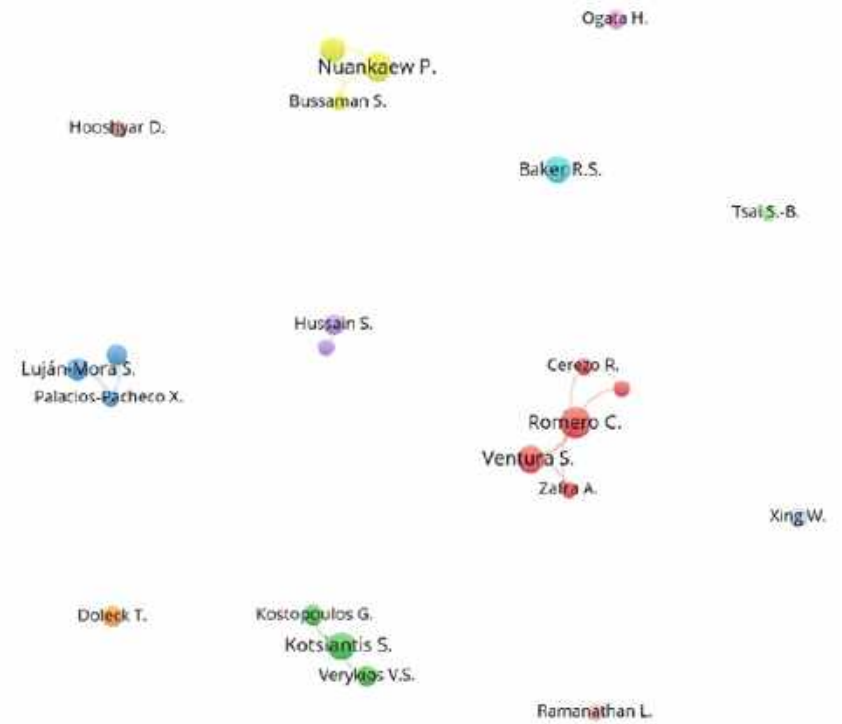

Figure 6. Collaboration between authors through co-authorship analysis (Source: Authors, using VOSviewer)

publications in the United States and India, with 170 and 166 publications, respectively. Spain had 85 publications, and Saudi Arabia had 61 publications.

Collaborative Zones of Literature

Co-authorship analysis by authors

Figure 6 shows the co-authorship analysis by authors for data mining in education, with 23 items grouped into 12 clusters. The minimum number of documents of an author threshold was set to five. Notably, five of the 12 clusters have multiple author collaboration. The author Romero C. has collaborated with Ventura S., Cerezo, R., Lara, J. A., and Zafra, A., and Romero, C. has a three-way collaboration with Ventura S. and Zafra A. as well as Cerezo R. and Lara J. A. The next three clusters have three authors each. The cluster in green shows that the authors Kostopouslos, G. and Verykios, V. S. have collaborated with Kotsiantis, S. individually, not with each other. The clusters in blue and yellow show three-way collaborations between the authors Luján-Mora, S., Palacios-Pacheco, X., and Villegas-Ch, W. and Bussaman, S., Nuankaew, P., and Nuankaew, W., respectively. The final collaboration is between the authors Hussain, S. and Salal, Y. K.

Co-authorship analysis by countries

Figure 7 shows the co-authorship analysis by countries, which depicts the collaborations between authors based on their affiliations. According to the analysis, China and the United States exhibit prominent author collaborations. The authors from China have been actively collaborating with authors from Australia and Thailand, whereas those from the United States have collaborated with Croatian and Canadian counterparts. Besides that, the green cluster highlights the active collaboration between authors from India and those from Bangladesh, Bulgaria, Finland, Greece, Iraq, Jordan, New Zealand, the Russian Federation, Serbia, and Turkey. Meanwhile, authors in Spain have been collaborating with authors from Belgium, Brazil, Chile, Ecuador, France, Italy, Peru, Poland, Portugal, Romania, and the United Kingdom, while in Saudi Arabia, the collaboration is with authors from Egypt, Pakistan, and Tunisia.

Co-citation analysis of cited references

The network visualization map in **Figure 8** depicts the co-citation relationship among cited papers with a 10-citation threshold. Post-threshold, 20 papers remain. The red and green clustered articles focus on data

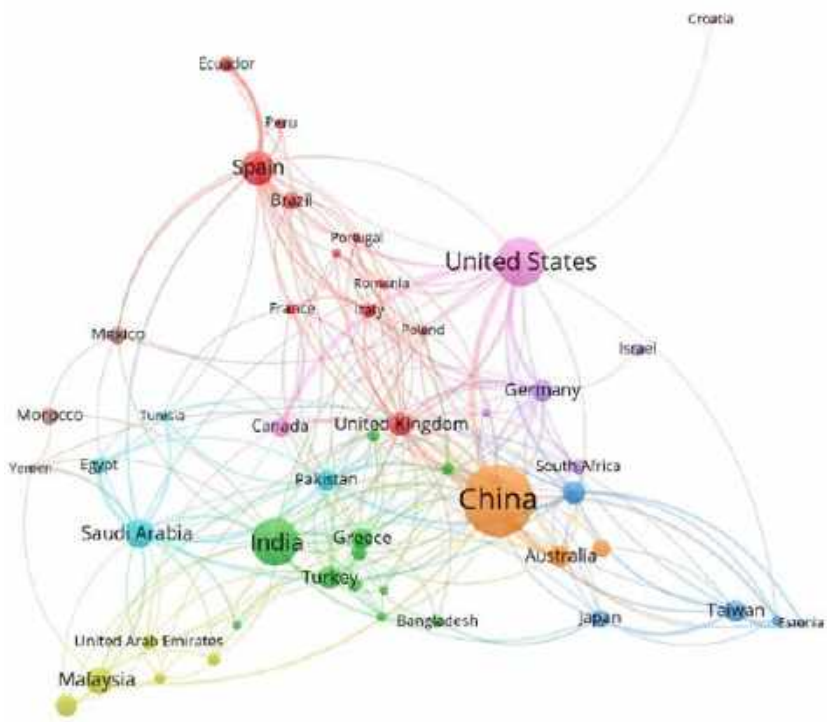

Figure 7. Collaboration between countries through co-authorship analysis (Source: Authors, using VOSviewer)

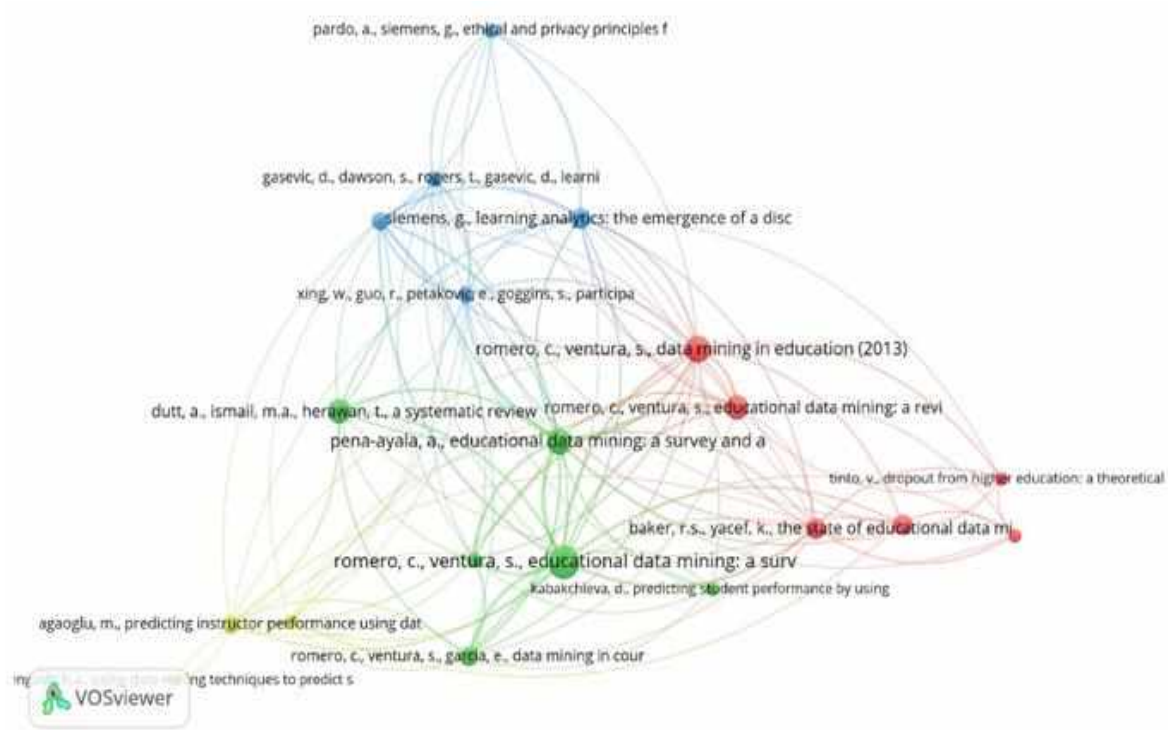

Figure 8. Relationship between cited papers through co-citation analysis (Source: Authors, using VOSviewer)

mining techniques, methods, and algorithms for educational data analysis. The articles in the blue cluster focus on learning analytics for educational enhancements.

The articles in the yellow cluster share a similar focus on predicting performances. However, the articles written by Mengash (2020) and Shahiri and Husain (2015) focus on predicting student performance, whereas Agaoglu (2016) focuses on predicting instructor performance. The author Romero, C.'s articles hold the top two spots for the highest co-cited journal articles published in 2007 and 2013, with 51 and 33 citations,

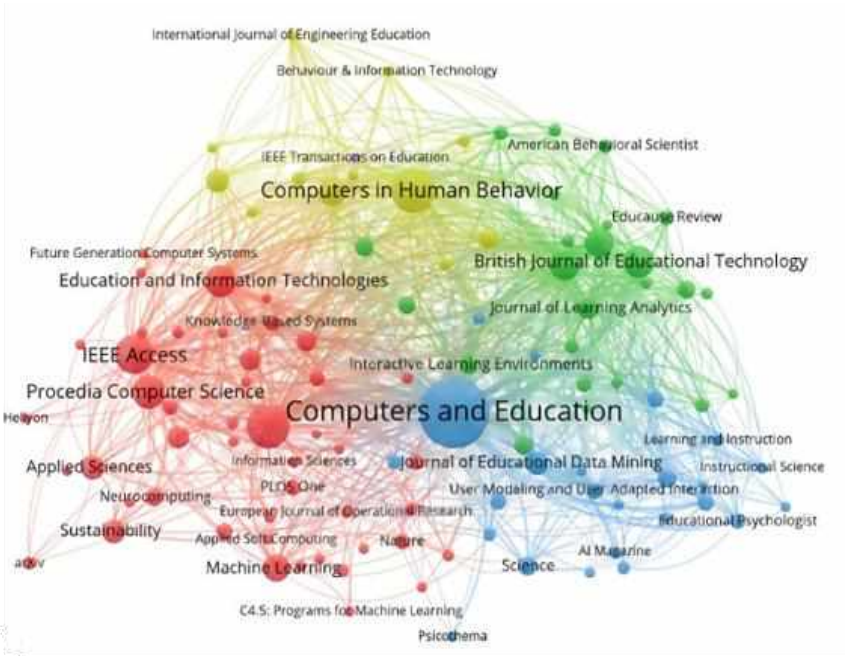

Figure 9. Relationship between cited sources through co-citation analysis (Source: Authors, using VOSviewer)

respectively. The former journal article is a survey, whereas the latter is a review of data mining in education, providing theoretical insights into data mining in education research.

Co-citation analysis of cited sources

Figure 9 shows the five clusters of co-cited sources obtained from the Scopus file analyzed in VOSviewer with a 30-citation threshold. The first cluster has 47 items, the second and third clusters have 24 items, respectively, the fourth cluster has 12 items, and the fifth cluster has only one item. The top-5 sources are "Computers and Education", "Computers in Human Behavior", "Expert Systems in Applications", "IEEE Access", and "Procedia Computer Science", with 1397, 552, 543, 424, and 308 citations, respectively. The red cluster journals focused on computer science, where data mining is a subfield. They collectively advance data mining techniques, algorithms, and applications across domains. Next, despite the green cluster emphasizing educational technology and learning analytics and the blue cluster emphasizing educational psychology and research, both clusters share the commonality of improving learning processes through technology, analytics, psychological factors, and research. On the other hand, the yellow cluster focuses on information technology and human-computer interaction, addressing usability and human implications in relation to technology.

Co-Occurrences of Keywords

Figure 10 shows a map of keywords using density visualization. The threshold used for the minimum number of occurrences of a keyword is five. The brighter the area surrounding a keyword, the more the occurrence of that keyword in the publications found.

Based on **Table 1**, which shows 16 of the most frequent keywords, it can be seen that apart from the keywords used in the search string, which include data mining, education, and data mining (highlighted in yellow), other keywords, such as students, prediction, academic performance, e-learning and learning system, machine learning and its algorithms such as decision trees and classification, learning analytics have high occurrences in the articles identified. Yagci (2022) exemplifies such research, using machine learning algorithms to predict students' academic performances in the final exam from prior grades. Besides that, data mining in education and learning analytics are keywords that are often used interchangeably, notably in higher education contexts (Aldowah et al., 2019). Aldowah et al. (2019) found that particular EDM and learning analytics techniques could be the most effective way to address some learning issues, tailoring strategies for student improvement. Next, with the recent events of the COVID-19 pandemic, studies focused on e-learning have also gained interest (Maatuk et al., 2021). Maatuk et al. (2021) highlight substantial use of e-learning platforms and systems during the pandemic by both teachers and students.

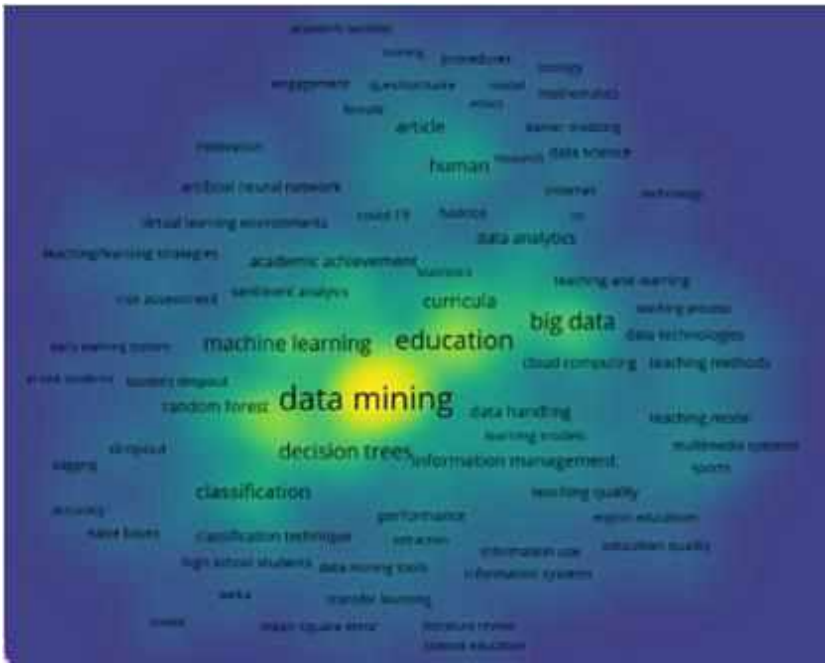

Figure 10. Map of keywords through co-occurrences analysis (Source: Authors, using VOSviewer)

Table 1. Top-16 of the most frequent keywords

Rank	Keywords	Occurrence	Total link strength
1	Data mining	951	5,240
2	Students	494	4,175
3	Education	422	3,316
4	Big data	263	1,480
5	Prediction	239	1,910
6	Academic performance	221	1,566
7	E-learning	206	1,683
8	Decision trees	180	1,343
9	Learning analytics	174	848
10	Learning systems	171	1,587
11	Teaching	156	1,360
12	Machine learning	154	1,003
13	Classification	141	909
14	Higher education	114	680
15	Educational institutions	96	864
16	Curricula	91	827

Impact & Performance of Articles & Citations in Field of Data Mining in Education

Overall citation metrics

Table 2 shows the overall citation metrics for all the final journal articles extracted from Scopus. The publication years are determined earlier in which the journal articles retrieved are from 2010 to 2022. However, despite the difference between the years 2010 to 2022 being 12 years, the citation years are 13 years instead. This is due to the data being extracted on April 16, 2023, which led to some articles being cited in 2023. The total number of papers extracted after pre-processing is 1,439 papers, amassing 24,455 citations overall. The average citation per year and citation per paper are 1881.15 and 16.99, respectively. The average number of authors per paper is 3.16. The h-index of the publications is 72, reflecting 72 papers with at least 72 citations each. The g-index reaches 111, indicating that the top 111 papers have collectively amassed over 12,321 citations.

Table 2. Overall citation metrics of finalized journal articles

Item	Data
Publication years	2010-2022
Citation years	13
Papers	1,439
Citations	24,455
Cites/year	1,881.15
Cites/paper	16.99
Authors/paper	3.16
h-index	72
g-index	111

Table 3. Top-5 highly productive authors

Rank	Author	TP	NCP	C	C/P	C/CP
1	Romero, C.	16	16	2100	131.25	131.25
2	Nuankaew, P.	14	11	103	7.36	9.36
3	Kotsiantis, S.	12	11	338	28.17	30.73
4	Ventura, S.	12	12	1944	162.00	162.00
5	Baker, R. S.	11	11	786	71.45	71.45

Note. TP: Total publications; NCP: Number of cited papers; C: Citations; C/P: Citations per paper; & C/CP: Citations per cited paper

Table 4. Top-5 highly productive countries

Rank	Country	TP	NCP	C	C/P	C/CP
1	China	366	273	2,714	7.42	9.94
2	United States	170	161	5,467	32.16	33.96
3	India	166	135	1,517	9.14	11.24
4	Spain	85	81	3,683	43.33	45.47
5	Saudi Arabia	61	58	1,163	19.07	20.05

Note. TP: Total publications; NCP: Number of cited papers; C: Citations; C/P: Citations per paper; & C/CP: Citations per cited paper

Impact & performance of authors

Table 3 shows the highly productive authors in data mining in education. All the publications by the author Romero, C. have had 2100 citations, with an average of 131.25 citations per paper. Next, out of 14 publications, 11 publications of the author Nuankaew, P. have been cited. The overall average citation and the cited publications are 7.36 and 9.36, respectively. The author Kotsiantis, S. has 12 publications, 11 of which are cited. The average citation of the publications and the cited publications are 28.17 and 30.73, respectively. The authors Ventura, S. and Baker, R. S. have 12 and 11 cited publications, respectively. The average citation per paper is 162.00 for Ventura, S. and 71.45 for Baker, R. S. Despite Romero, C.'s greater publication count, Ventura S.'s papers have received more citations per publication.

Impact & performance of countries

The highly active countries in terms of publications are shown in **Table 4**. Every country listed above has a total number of cited publications less than the number of total publications. China has the highest number of publications, with 2,714 citations, an average citation per paper, and an average citation per cited paper of 7.42 and 9.94, respectively. However, the total citations of the cited publications of the United States and Spain are higher than those of China. The citation per cited paper for the United States is 33.96. Meanwhile, Spain attains 45.47. India's 1517 citations span 135 cited papers, with an average of 11.24 citations per cited paper. Saudi Arabia has 58, gathering 1163 citations, averaging 19.07 citations per paper and 20.05 citations per cited paper.

Impact & performance of journals

Table 5 outlines the top-10 journals leading data mining in education research. Notably, the "International Journal of Emerging Technologies in Learning" stands out with 62 publications with 981 citations for 61 papers. The average citation per paper is 15.82, and the average citation per cited paper is 16.08.

Table 5. Top-10 highly productive journals

Rank	Source title	TP	NCP	C	C/P	C/CP
1	International Journal of Emerging Technologies in Learning	62	61	981	15.82	16.08
2	Education and Information Technologies	39	38	846	21.69	22.26
3	IEEE Access	38	37	927	24.39	25.05
4	International Journal of Advanced Computer Science & Applications	35	31	251	7.17	8.10
5	Applied Sciences Switzerland	27	23	398	14.74	17.30
6	Sustainability Switzerland	23	21	295	12.83	14.05
7	Boletin Tecnico Technical Bulletin	19	14	35	1.84	2.50
8	Computers and Education	19	19	2,107	110.89	110.89
9	Computer Applications in Engineering Education	17	17	504	29.65	29.65
10	Computers in Human Behavior	17	17	1436	84.47	84.47

Note. TP: Total publications; NCP: Number of cited papers; C: Citations; C/P: Citations per paper; & C/CP: Citations per cited paper

Table 6. Top-10 highly impactful journals articles

No	Reference	Title	Source	TC	C/Y
1	Romero and Ventura (2013)	Data mining in education	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery	552	55.20
2	Romero et al. (2013)	Predicting students' final performance from participation in online discussion forums	Computers and Education	416	41.60
3	Daniel (2015)	Big data and analytics in higher education: Opportunities and challenges	British Journal of Educational Technology	329	41.13
4	Asif et al. (2017)	Analyzing undergraduate students' performance using educational data mining	Computers and Education	323	53.83
5	Romero and Ventura (2020)	Educational data mining and learning analytics: An updated survey	Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery	295	98.33
6	Costa et al. (2017)	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses	Computers in Human Behavior	249	41.50
7	Romero et al. (2013)	Web usage mining for predicting final marks of students that use Moodle courses	Computer Applications in Engineering Education	203	20.30
8	Shapiro et al. (2017)	Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers	Computers and Education	194	32.33
9	Cerezo et al. (2016)	Students' LMS interaction patterns and their relationship with achievement: A case study in higher education	Computers and Education	194	27.71
10	Hu et al. (2014)	Developing early warning systems to predict students' online learning performance	Computers in Human Behavior	191	21.22

Note. TP: Total publications; NCP: Number of cited papers; C: Citations; C/P: Citations per paper; & C/CP: Citations per cited paper

As seen in **Figure 10**, "Computers and Education" has the most citations, akin to the largest node. The cited papers are equal to number of publications in journal, and average citation per paper is 110.89. "Computers in Human Behavior" follows with 1,436 citations and 84.47 citations per paper among 17 publications.

Impact & performance of journal articles

The top-10 articles with the highest number of citations are presented in **Table 6**. It can be seen that four out of the ten articles have Romero, C. and Ventura, S. as the collaborative authors in the publications. The most cited article is "Data mining in education", published in 2013 and co-authored by Romero, C. and Ventura, S., averaging 55.2 citations annually. However, the article co-authored by the same authors titled "Educational data mining and learning analytics: An updated survey" has the highest number of citations per year among these ten articles, which is 98.33. Both articles are from "Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery" journal.

Nevertheless, as shown in **Table 5**, the journal "Computers and Education" has the highest number of citations per publication, which is reflected in **Table 6** as well. The "Computers and Education" journal has

four articles in the top-10 articles listed above. The articles are "Predicting students' final performance from participation in online discussion forums", "Analyzing undergraduate students' performance using educational data mining", "Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers", and "Students' LMS interaction patterns and their relationship with achievement: A case study in higher education" with 41.6, 53.83, 32.33, and 27.71 citations per year, respectively.

Findings Comparison of This Study With that of Wang et al. (2022)

Table 7 provides the similarities and differences of BA findings of this study and Wang et al. (2022).

Table 7. Similarities & differences of this study & Wang et al. (2022)

Item	Similarities	Differences
Total publications	None	Total number of publications in WoS obtained by Wang et al. (2022) is 334 papers, whereas total number of publications extracted in this study through Scopus is 1,439 papers after pre-processing.
Publication by years	Both studies have an increasing distribution of publications from 2010 to 2020 & between 2021 & 2022.	This study has a slightly lower number of publications in 2021 compared to 2020, whereas Wang et al. (2022) found that publications in WoS have increased consistently after 2020.
Authors	Romero, C. is leading author in terms of publications in WoS, with seven publications according to Wang et al. (2022) & 16 publications in Scopus on this study. Only Ventura, S. & Kotsiantis, S. after Romero, C. are in top-5 highly productive authors in both databases.	Other two authors in top-5 are Jones, K. M. L. & Zhang, W. with five & four publications in WoS (Wang et al., 2022) & Nuankaew, P. & Baker, R. S. with 14 & 11 publications in Scopus.
Journals	Similar journals that are in top-10 journals in field of data mining in education in both WoS according to Wang et al. (2022) & Scopus in this study are Education & Information Technologies, IEEE Access, Sustainability, Computer Applications in Engineering Education, & Computers & Education. In terms of citation, Computers & Education journal has highest number of citations in WoS according to Wang et al. (2022) & Scopus in this study with 722 & 2,107 citations, respectively.	Journal with highest number of publications in WoS, according to Wang et al. (2022), is IEEE Access with 26 publications. But journal in first rank in Scopus in this study is International Journal of Emerging Technologies in Learning, with 62 publications.
Countries	In terms of countries, China holds top-spot for greatest number of publications in both databases, with 97 & 366 publications, respectively, in WoS according to Wang et al. (2022) & Scopus in this study. Spain, India, & the USA are also in top-5 countries in both studies. The USA ranks second for highest number of publications, with 170 & 57 publications in Scopus & WoS, respectively.	India is in third rank with 166 publications in this study, whereas in Wang et al. (2022), India is in fourth rank with 17 publications. Meanwhile, Spain is in third rank with 35 publications in Wang et al. (2022), whereas in this study Spain is in fourth rank with 85 publications. In Wang et al. (2022), fifth country with highest number of publications is Turkey, but in this study, fifth country with highest number of publications is Saudi Arabia.
Keywords	Keyword with highest frequency in WoS on findings in Wang et al. (2022) & Scopus is data mining, with frequencies of 158 and 951, respectively. Similar keywords found in 16 most frequent keywords in both studies are data mining, students, education, big data, prediction, academic performance, e-learning, learning analytics, machine learning, classification, & higher education.	Different keywords that are in 16 most frequent keywords in Wang et al. (2022) are model, online, analytics, MOOC, & data analytics. In Scopus, keywords are decision trees, learning systems, teaching, educational institutions, & curricula.
Articles	Out of top-5 most important articles identified by Wang et al. (2022) & top-10 most important publications found in this study, only one common article was found, which is "Predicting students' final performance from participation in online discussion forums" published in 2013 & written by Romero et al. (2013). Article gained 416 citations in Scopus & 39 citations in WoS. Articles that rank first in WoS & Scopus are both co-authored by Romero, C. & Ventura, S. But both articles are not same.	Article that got highest number of citations in WoS is "Educational data mining: A review of the state of the art," which was published in 2010 & has a total of 118 citations. Meanwhile, article with highest number of citations in Scopus is "Data mining in education". This article was published in 2013 & gained a total of 552 citations.

Table 7 (Continued). Similarities & differences of this study & Wang et al. (2022)

Item	Similarities	Differences
Cited references	None	Cited references in Wang et al. (2022) consist of three clusters focusing on educational technology, computer technology & algorithms, & review studies. Cited references in this study consist of four clusters. Two clusters focus on data mining techniques, methods, & algorithms, one focuses on learning analytics & final cluster focuses on performance prediction.

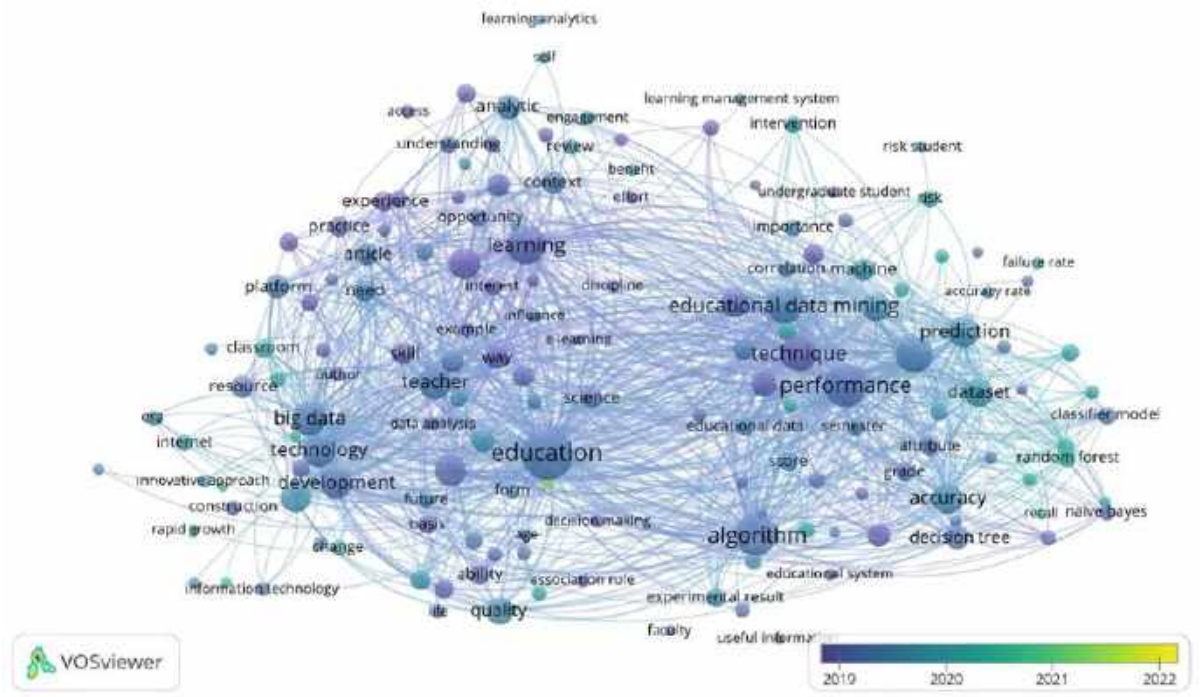

Figure 11. Terms map analysis (Source: Authors, using VOSviewer)

Evolution of Research Topics

Figure 11 shows the evolution of research topics in the field of data mining in education. The map starts in the year 2019 and goes on up till the year 2022. The blue or cool color themes indicate the older research topics, whereas the yellow or hot color themes indicate the newer research topics foci. There is only one yellow node for the term "COVID-19". This indicates that COVID-19-related research topics emerged in 2022. Research with specific data mining techniques and machine learning was the focus in 2020 and 2021. Other research topics, such as those related to academic performance predictions and online learning, have garnered attention since 2019.

DISCUSSION

Key Findings

The author Romero, C. is the most prolific in the field of data mining in education. He has 16 publications (based on **Figure 3**) available in Scopus, where two of his articles, which are co-authored with other authors, i.e., López, M.-I., Luna, J.-M., and Ventura, S., are in the top two spots for the most impactful and recognized publications in the field of data mining in education as shown in **Table 6**.

Yet, another article co-authored by Romero, C. and Ventura, S. has the highest number of citations per year, titled "Educational data mining and learning analytics: An updated survey". Hence, alongside Romero, C. it can be inferred that Ventura, S. also significantly impacts this field. To support this, Ventura, S. has the

second highest total citations for all his publications, which is 1,944, after Romero, C., who has 2,100 citations for all his publications, as shown in **Table 3**. Both authors are also from the same institution, University of Cordoba in Spain, which has recorded the highest number of publications, as shown in **Figure 3**.

Next, despite China having the greatest number of publications with the largest node in **Figure 5**, the United States tops as the country with the most influential publications, followed by Spain with 5,467 citations and 3,683 citations, respectively. However, papers written by authors in Spain have higher citations per cited paper. This indicates that the cited papers published in Spain have made substantial contributions to the field of data mining and education.

Besides that, according to the findings in **Table 5**, the journal "Computers and Education" has the highest total number of citations for all its publications, which is 2,107 citations with 110.90 citations per cited paper, followed by the journal "Computers in Human Behavior" with a total of 1,436 citations with 84.47 citations per cited paper. Four articles published in the "Computers and Education" journal and two in the "Computers in Human Behavior" journal are in the top-10 list of the most impactful articles in this field. All six articles also have total citations within the range of 191 to 416 (see **Table 6**). The high number of citations indicates the value of the articles published in this journal for this particular field of research. This suggests that researchers may find that the articles in these journals help shape subsequent research directions.

In terms of the keywords, the findings in **Table 1** and **Figure 10** consistently suggest that the articles that have been published indicate that the research focuses revolve around the use of educational data to predict students' academic performance through machine learning algorithms and learning analytics. Much work related to e-learning also has gained additional attention during the COVID-19 pandemic. These keywords suggest that authors in this field actively research to improve students' academic outcomes and pedagogical strategies, resolve issues related to e-learning, and provide insight into the aspects that need to be considered due to COVID-19. The findings of the keywords in **Table 1** align with the evolution of research topics, as shown in **Figure 11**, where these aspects have gained researchers' interest since 2019.

The co-citation of cited references indicates that articles from the same research area tend to cite each other. For example, three identified clusters for co-citation of cited references are identified, as shown in **Figure 8**. These can be categorized into data mining methods and algorithms that are used to gain insight and useful information from educational data, learning analytics for education-related improvements, and studies that revolve around predicting the performances of educators and learners. The most commonly cited articles from each cluster provide the basic knowledge for subsequent research. For instance, the paper by Romero, C. and Ventura, S. titled "Educational data mining: A survey from 1995 to 2005" in the green cluster is the largest node. The article surveys the use of data mining techniques in different educational systems, such as traditional educational systems, web-based courses, learning content management systems, and adaptive and intelligent web-based educational systems (Romero & Ventura, 2007). The findings of this study act as a guide for the advancement of research in the field of data mining in education in the context of educational systems.

Finally, regarding the comparison between this study and that of Wang et al. (2022), the main difference is in the number of publications, where there is a huge gap of 1,105 publications, which caused some differences in the findings of both studies. This suggests that most authors have more publications in Scopus than WoS. However, this does not mean that the articles in Scopus outweigh those in WoS. There are several articles identified, as shown in **Table 7**, which are only available in either one of the databases. Hence, it can be inferred that both databases have equal importance in providing the academic resources needed for subsequent research in data mining in education.

Current Research Areas & Authors for Collaboration

The articles Esteban et al. (2021) and López-Zambrano et al. (2021, 2022) are recent articles that author Romero, C. has co-authored. All three articles focus on predicting students' academic performance (Esteban et al., 2021; López-Zambrano et al., 2021, 2022). According to López-Zambrano et al. (2021), more of these focus areas can be researched. For example, researchers may conduct further research on predicting student performance in primary education, as very little research is conducted in this sector.

Not only that, the author Nuankaew, P. also has two recent publications, Nuankaew and Nuankaew (2022) and Nuankaew et al. (2021), which also focused on predicting student performance and achievements using different models. Nuankaew and Nuankaew (2022) highlighted that future research should focus on utilizing different machine learning models to improve student performance prediction and provide better learning experiences for each learner. Moreover, the study conducted by Yacoub et al. (2022), co-authored by Ventura, S., also focused on using machine learning to predict students' performance. The study highlights the importance of proper academic performance prediction in assisting educational institutions to take proper measures and provide students with the right support to ensure the betterment of their academic performances (Yacoub et al., 2022).

Besides that, the author Kotsiantis, S. has also co-authored and contributed to this research area. Kotsiantis, S. has 10 out of 12 publications on predicting student performance using different models or machine learning algorithms. For instance, the studies Alachiotis et al. (2022) and Tsiakmaki et al. (2021) use fuzzy logic-based automated machine learning and supervised machine learning methods to predict student performance, respectively. Alachiotis et al. (2022) also showed that through a voting generalization procedure involving three of the most accurate classifiers and the default parameters of learning algorithms, the prediction outcome is much higher and more accurate than only using a single-tuned learning algorithm.

Next, another common and recent research area identified is e-learning. Especially since the COVID-19 pandemic, e-learning has become the primary source for obtaining big educational data (Wang et al., 2022). With the many interactions of learners through LMS, researchers have been able to use these large amounts of data to understand and evaluate educational processes in conjunction with making decisions to improve the effectiveness of the current education system (Fischer et al., 2020). A study conducted by Nuankaew and Nuankaew (2021) suggests that e-learning and the traditional settings of the educational context contribute equally to students' academic achievements based on the data collected before and during the COVID-19 pandemic. Hence, this analysis concludes that the prevalent research areas that researchers can focus on for subsequent research may revolve around predicting student performances using machine learning algorithms and data mining in e-learning. Researchers may collaborate with prominent authors such as Romero, C., Ventura, S., Nuankaew, P., Kotsiantis, S., and Baker, R. S., as these authors have been in data mining in education for a long time and have consistently published research articles from 2010 to 2022. They may offer more insights and knowledge regarding the field for newer research topics through collaborations.

It is anticipated that several key developments are emerging in the continually evolving landscape of data mining in education. The substantial advancements in machine learning for performance prediction and e-learning revealed by BA will catalyze the research and development of adaptive learning environments that incorporate data mining to change course materials in real time based on student engagement, characteristics, and performance to create personalized learning. Future work will also focus on automated feedback systems that use real-time analytics to provide quick, data-driven feedback to teachers and students. Multi-modal analytics is another trend that seeks to build a more complete picture of the learning process by incorporating several data streams such as eye-tracking, keyboard dynamics, and physiological markers. There is also an urgent need for more extensive and varied datasets alongside these developments, particularly in the K-12 sector. In addition, the Industrial Revolution 4.0 also sparked a surge in data mining approaches in fields like IoT, to produce more integrated, smarter learning environments. As importantly, addressing the ethical and privacy issues related to collecting and analyzing educational data will become more crucial as the field develops.

CONCLUSIONS

In conclusion, BA offers valuable insights into research trends within a specific field. This study has effectively identified current research trends and assessed the impact and performance of publications in the field of data mining in education synonymous with big educational data, aiding subsequent research directions. It highlights potential collaborators and journals for enhanced visibility and impact. By analyzing Scopus data, this research contributes to an additional bibliometric analysis in this field, complementing Wang et al. (2022). As a summary of the findings, China maintains its leading position in EDM publications, while prolific author Romero from the University of Cordoba emerges as a significant contributor. Spain's influence

is equally notable in this domain. China and the United States are countries with the highest levels of author collaborations, and "Computers and Education" stands out with the highest citation count, solidifying its impact as the premier journal in data mining and big data in education. Comparison with Wang et al. (2022) unveils differences due to the significant difference in the total number of publications found in Scopus compared to WoS. Nevertheless, parallels in trends offer insights.

Limitations of this study include the exclusion of highly cited materials not available in Scopus, potentially impacting findings. Notably, factors like the h-index and g-index of each article remain unexplored, affecting insight into article impact and performance. Future researchers may consider utilizing untapped databases such as ERIC and PsycINFO for bibliometric study on data mining in education. While BA offers descriptive insights, a mixed-method approach, such as an SLR with bibliometric analysis, will provide a richer and more detailed understanding of data mining in education. To sum up, the field of data mining in education has shown tremendous advancement since the year 2010 and 12 years later. It is definitely a field that is still growing and has more to offer in terms of research to ensure that the education sector continues to thrive in enhancing the educational experience.

Author contributions: Both authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. Both authors approved the final version of the article.

Funding: This study was supported by the Malaysian Ministry of Higher Education, Fundamental Research Grant Scheme, FRGS/1/2020/SS10/UNIMAS/01/1.

Ethics declaration: The authors declared that study did not require ethics committee approval as it is based on existing literature. The study does not involve human participants.

Declaration of interest: The authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. *IEEE Access, 4*, 2379-2387. https://doi.org/10.1109/ACCESS.2016.2568756

Ahmi, A. (2021). *Bibliometric analysis for beginners: A starter guide to begin with a bibliometric study using Scopus dataset and tools such as Microsoft Excel, Harzing's Publish or Perish and VOSviewer software*. Aidi-Ahmi.

Alachiotis, N. S., Kotsiantis, S., Sakkopoulos, E., & Verykios, V. S. (2022). Supervised machine learning models for student performance prediction. *Intelligent Decision Technologies, 16*(1). 93-106. https://doi.org/10.3233/IDT-210251

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37*, 13-49. https://doi.org/10.1016/j.tele.2019.01.007

Baek, C., & Doleck, T. (2022). Educational data mining: A bibliometric analysis of an emerging field. *IEEE Access, 10*, 31289-31296. https://doi.org/10.1109/ACCESS.2022.3160457

Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Eds.), *Cambridge handbook of the learning sciences* (pp. 253-274). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.016

Esteban, A., Romero, C., & Zafra, A. (2021). Assignments as influential factor to improve the prediction of student performance in online courses. *Applied Sciences, 11*(21). 10145. https://doi.org/10.3390/app112110145

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education, 44*(1), 130-160. https://doi.org/10.3102/0091732X20903304

Hermaliani, E. H., Fanani, A. Z., Santoso, H. A., Affandy, A., Purwanto, P., Muljono, M., Syukur, A., Setiadi, D. R. I. M., & Rafrastara, F. A. (2022). Systematic review of educational data mining for student performance prediction using bibliometric network analysis (SeBriNA). In *Proceedings of the 2022 International Seminar on Application for Technology of Information and Communication* (pp. 463-468). https://doi.org/10.1109/iSemantic55962.2022.9920477

International Educational Data Mining Society. (2011). *EDM*. https://educationaldatamining.org/

López-Zambrano, J., Lara, J. A., & Romero, C. (2022). Improving the portability of predicting students' performance models by using ontologies. *Journal of Computing in Higher Education, 34*, 1-19. https://doi.org/10.1007/s12528-021-09273-3

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema, 33*(3), 456-465. https://doi.org/10.7334/psicothema2021.62

Maatuk, A. M., Elberkawi, E. K., Aljawarneh, S., Rashaideh, H., & Alharbi, H. (2022). The COVID-19 pandemic and e-learning: Challenges and opportunities from the perspective of students and instructors. *Journal of Computing in Higher Education, 34*, 21-38. https://doi.org/10.1007/s12528-021-09274-2

Marín-Marín, J., López-Belmonte, J., Fernández-Campoy, J., & Romero-Rodríguez, J. (2019). Big data in education. A bibliometric review. *Journal of Social Sciences, 8*(8), 223. https://doi.org/10.3390/socsci8080223

Masood, M., & Mokmin, N. A. M. (2017). Case-based reasoning intelligent tutoring system: An application of big data and IoT. In *Proceedings of the 2017 International Conference on Big Data Research* (pp. 28-32). https://doi.org/10.1145/3152723.3152735

Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access, 8*, 55462-55470. https://doi.org/10.1109/ACCESS.2020.2981905

Menon, A., Gaglani, S., Haynes, M. R., & Tackett, S. (2017). Using "big data" to guide implementation of a web and mobile adaptive learning platform for medical students. *Medical Teacher, 39*(9), 975-980. https://doi.org/10.1080/0142159X.2017.1324949

Nuankaew, P., & Nuankaew, W. S. (2022). Student performance prediction model for predicting academic achievement of high school students. *European Journal of Educational Research, 11*(2), 949-963. https://doi.org/10.12973/EU-JER.11.2.949

Nuankaew, P., Nasa-ngium, P., & Nuankaew, W. S. (2021). Application for identifying students achievement prediction model in tertiary education: Learning strategies for lifelong learning. *International Journal of Interactive Mobile Technologies, 15*(22), 22-43. https://doi.org/10.3991/IJIM.V15I22.24069

Rodrigues, M. W., Isotani, S., & Zárate, L. E. (2018). Educational data mining: A review of evaluation process in the e-learning. *Telematics and Informatics, 35*, 1701-1717. https://doi.org/10.1016/j.tele.2018.04.015

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*, 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 40*(6), 601-625. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(1), e1187. https://doi.org/10.1002/widm.1187

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*. 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics–A literature review. *ICTACT Journal on Soft Computing, 4*(5), 1035-1049. https://doi.org/10.21917/ijsc.2015.0145

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2021). Fuzzy-based active learning for predicting student academic performance using autoML: A step-wise approach. *Journal of Computing in Higher Education, 33*(3), 635-667. https://doi.org/10.1007/s12528-021-09279-x

Wang, C., Dai, J., & Xu, L. (2022). Big data and data mining in education: A bibliometrics study from 2010 to 2022. In *Proceedings of the 7th International Conference on Cloud Computing and Big Data Analytics* (pp. 507-512). https://doi.org/10.1109/ICCCBDA55098.2022.9778874

Yacoub, M. F., Maghawry, H. A., Helal, N. A., Gharib, T. F., & Ventura, S. (2022). An enhanced predictive approach for students' performance. *International Journal of Advanced Computer Science and Applications, 13*(4), 879-883. https://doi.org/10.14569/IJACSA.2022.01304101

Yagci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*, 11. https://doi.org/10.1186/s40561-022-00192-z