Research Article

Machine Learning-based Predictive Systems in Higher Education: A Bibliometric Analysis

Fati Tahiru[1,*], Steven Parbanath[1], Samuel Agbesi[2]

[1]Department of Information Technology, Durban University of Technology, KwaZulu Natal, SOUTH AFRICA.
[2]Department of Computer Science, IT University of Copenhagen, Copenhagen, DENMARK.

ABSTRACT

This paper aims to comprehensively review the present state and research trends in predictive systems in higher education. It also addresses the research contribution of countries in Machine Learning-based predictive systems in higher education to depict the research landscape given the growing number of related publications. A bibliometric analysis of publications on predictive systems in education published in the Scopus Database from 2015 to 2022 was conducted. The dataset obtained covered the contribution of authors, affiliations, countries, themes and trends in the field of Machine Learning-based predictive systems in higher education. A total of 72 publications with 3408 cited references were collected from Scopus for the bibliometric analysis. The technique used for the bibliometric analysis included performance analysis and science mapping. Research on Machine Learning-based predictive systems has been widely published from 2020 to 2022. Researchers in China, Belgium, Spain, India, and Korea were most active in researching Machine Learning-based predictive systems in education. However, international collaborations have remained infrequent except for the few involving Australia, Belgium, and Canada. There is a lack of research in the subject area in Africa. This study illustrates the intellectual landscape of Machine Learning-based predictive systems in higher education and the field's evolution and emerging trends. The findings highlight the area of research concentration and the most recent developments and suggest future research collaborations on a larger scale as well as additional research on the implementation of predictive systems in education in Africa.

Keywords: Machine Learning, Predictive System, Data Analytics, Learning Analytics, Higher Education.

Correspondence:
Fati Tahiru
Department of Information Technology,
Durban University of Technology,
SOUTH AFRICA.
Email: 22176488@dut4life.ac.za
ORCID ID: 0000-0003-0874-0428

Received: 18-11-2022;
Revised: 29-05-2023;
Accepted: 23-06-2023.

INTRODUCTION

Machine Learning (ML) is a branch of artificial intelligence and computer science that uses data and algorithms to mimic human learning. The algorithm used in ML is not explicitly programmed but learns from data and experience,[1] and its use has become pervasive. Several industries that collect, store, and interact with data have adopted ML. Amazon and Microsoft, for instance, have ML agents like Alexa and Cortana that interact with users. Amazon's Alexa, for example, allows users to perform a multitude of tasks, such as turning on music, television, light and many more.

Similarly, Microsoft's Cortana can assist users in opening Apps on their computers and carrying out specific tasks, such as setting reminders and alarms for meetings and events.[2] ML tools and platforms have also enormously benefited the education sector.[3]

DOI: 10.5530/jscires.12.2.040

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

Machine learning applications in higher education are advancing rapidly, and some use cases in education include systems for predicting student retention rates and providing personalised feedback to students. For instance, ML has been successfully used to predict students at risk of dropping out of school.[4]

Numerous studies[1,5,6] on ML-based predictive systems in higher education have employed a Systematic Literature Review (SLR) methodological approach. However, none have focused on the research clusters that could address the topic areas requiring additional research.

However, few studies employ bibliometric analysis to shed light on the applications of ML in higher education. This suggests a promising field in bibliometric analysis that requires additional research to provide researchers and educational stakeholders with insight and direction on current and future trends in predictive systems in higher education.

Bibliometric analysis is an excellent method for the research questions presented in this study as it highlights the qualities of publications in a particular direction.[7] The technique provides high-level insights into the characteristics of many publications within a particular research domain.[8] Bibliometrics is a subfield of information science that focuses on the intersection and synthesis of fields such as information science, mathematics, and science. The Bibliometrics methodology has been utilised to provide insight into a variety of disciplines such as process safety,[8] health care,[7,9,10] blockchain and supply chain,[11,12] E-learning,[13] Artificial intelligence,[14,15] tourism,[16] Smart education[17,18] and many more. To the best of our knowledge, no research has been conducted on ML-based predictive systems in education using Bibliometric analysis. In order to achieve the objectives of this study, bibliometric analysis is employed to gain insight and comprehension into the research themes and trends that require further study in ML-based predictive systems in higher education.

This paper aims to highlight the research conducted on ML-based predictive systems in higher education from 2015 to 2022. The paper presents a comprehensive bibliometric review focusing on the patterns and trends concerning collaborations, key publications, countries' contributions, significant topics, and trends in the research field. The study specifically addresses the following research questions:

RQ1. Who are the most productive and influential authors and institutions in the field of Machine learning-based predictive systems in higher education?

RQ2. What are the main themes and trends found in the analysed topic?

RQ3. What are the collaboration patterns and research contributions among countries in the analysed topic?

The remaining chapters are organised in the following order; section 1 presents an introduction to the study, section 2 the related works, and Section 3 describes the procedure for data collection and screening methods used. Section 4 presents the results of the analysis, providing answers to the above-listed research questions, followed by discussions of significant findings and future directions for researchers and practitioners. Finally, Section 5 presents the conclusion.

Related Works

This section presents the ML categories and their implementations in education. Subsequently, the section discusses studies utilising the SLR and Bibliometric Analysis and presents the identified research gap.

ML categories and their implementations in education

Predictive systems in higher education have gained much popularity over the years due to their ability to identify at-risk students early and implement early measures to retain them.[19] The three main categories of ML generally described in the field are Supervised Learning, Unsupervised Learning and Reinforced Learning. Supervised learning involves feature labels for training and predicting future events;[1] Unsupervised Learning does not use label data, whereas Reinforced learning uses the reward system to train the algorithm.[20]

Predictive Systems in higher education have utilised these three categories of ML methods to improve teaching and learning in the educational sector. For example, Yousafzai, Hayat and Afzal[21] developed a ML system to predict students' scores and grades based on their academic performance in the past. The study applied Supervised ML techniques for Classification and Regression. Classification systems predicted grades, whereas Regression techniques predicted the marks of students. In another study, Amalina Diyana Suhaimi *et al.*[19] compared three Supervised ML algorithms, Logistic Regression, Support Vector Machine, and k-Nearest Neighbor, to investigate Classification approaches for resolving the issue of learner retention at Open University Malaysia. Also, Sassirekha and Vijayalakshmi[22] developed a new hybrid methodology based on the best-supervised model that can consistently forecast the output factor of student performance. The model assisted students in improving their academic performance by providing performance forecasts via its monitoring system each semester to guarantee that all students pass exams on their first attempt.

Other studies have utilised Unsupervised ML methods in building predictive systems. For example, Liu[23] used the Deep Denoising autoencoder, a standard Neural Network model, to evaluate the teaching quality of a university. Likewise, Kehinde *et al.*[24] used the Artificial Neural Network, an Unsupervised ML model, to predict student success.

On the other hand, Liu and Ardakani[25] used the Reinforced Learning approach to analyse learners' emotional states and instantly recommend the most appropriate content to keep students in a good mood. The use of Reinforced Learning has also been utilised in language learning to assist students in learning the Kazakh Latin alphabet through a chatbot.[26]

Studies that utilised the Systematic Literature Review and Bibliometric Analysis

Some Systematic Literature Review studies have been conducted on ML-based predictive systems in higher education. The systematic Literature review methodology efficiently systematises knowledge in a particular research domain. Numerous articles have provided analyses of various topics on the predictive system. For example, De Oliveira, Bernardini and Viterbo[5] conducted a SLR to identify the use of Educational Data Mining tools or techniques for implementing Recommendation Systems, focusing on preventing student retention in higher education programs. The study determined what methods of ML were used in developing Recommendation Systems in the context of Educational Data Mining and identified the most used methods.

Thirty-one (31) articles were selected from five (5) Databases to conduct the literature review. The study by Umer *et al.*[6] aimed at how predictive analytics have been applied in the higher education domain. The authors explore the current trends in building data-driven predictive models to determine student performance and the ML techniques and strategies used in previous studies to develop predictive systems. The literature relevant to the study spans from 2008 to 2018. Similarly, Sandra, Lumbangaol and Matsuo[1] examined the extent of research on implementing ML algorithms and modelling student performance. The study's search for the literature review spanned from 2019 to 2021 with 11 articles.

Studies that utilised Bibliometric Analysis

Particularly relevant to our study is Bibliometric Analysis. There have been some studies on Bibliometric Analysis in AI applications and Smart education. For example, Shukla, Muhuri and Abraham[14] presented a bibliometric study on fuzzy techniques in big data. The authors conducted a bibliometric coupling between authors and countries using articles in the Web of Science and Scopus databases. The results demonstrated the research clusters in publications based on authors, institutions and countries.

Similarly, Zhao *et al.*[15] conducted a large-scale Bibliometric Analysis and Systematic Review of the trends in applying Artificial Intelligence technology to wastewater treatment, with a literature search in the Science Citation Index from 1995 to 2019. The analysis based on a co-word network was used to identify the topics and trends in applying AI in wastewater treatment. Likewise, in the field of smart education, Guo, Li and Guo[17] conducted a bibliometric analysis focusing on co-authorship between authors, countries, regions and organisations to present a smart education integrated knowledge structure map for research direction. The study made use of 2358 Web of Science articles from the period 2000 to 2021. The findings on countries' collaboration in the research area indicate that the USA has the largest number of publications and citations in smart education. The study also revealed that most countries involved in smart education research focused on theoretical and practical research of smart education. The co-keyword analysis identified the new research spot as education, technology, students, higher education, impact, performance and internet.

Similarly, Li and Wong,[18] presented a comprehensive bibliometric review on the current landscape of smart education. The authors analysed the study on 1317 publications from Web of Science and Scopus databases. The findings of the most active research collaboration and the cited references were consistent with the findings of Guo. However, the emerging topics identified were the Internet of Things, big data, flipped learning and gamification.

Summary and research gap

The rationale of predictive systems in higher education has been well established in SLR methodology.[1,5,6] The research approaches based on SLR provide information about authors and publications in predictive systems in education; however, the analysis and synthesis do not provide a comprehensive understanding of the structure of the research field. Also, the analysis lacks clusters and narrations on research trends. In addition, SLR fails to evaluate the contributions of various countries in the research area. Some important and useful information, such as collaboration between authors, universities and countries in the field, most influential authors, universities, and countries, research topics and trends being pursued by researchers in the field are not answered in traditional review methods.[8] Few studies[18,19] have used a bibliometric approach to examine the relationships between authors, affiliations, and institutions in the field of ML applications in higher education. However, these studies did not seek to facilitate international research collaboration or clarify thematic clusters to guide future research. Although some progress has been made in bibliometric analysis in higher education, implementing a ML-based predictive system for higher education requires clarity in research directions. We contribute to the landscape of ML and higher education by conducting one of the most comprehensive bibliometric reviews of this emerging important topic in the field of ML-based predictive systems in higher education to address future research trends.

METHODOLOGY

This section explains how the search for scientific documents was conducted and also presents how the bibliometric analysis technique is applied in this study.

The data was gathered from only the Scopus database to manage all publication metadata consistently and standardised. Typically, Bibliometric studies can employ a single database, which is also practically required and acceptable.[27] This database is also popular among the databases with highly indexed impact. The search keywords such as "predictive systems"," machine learning", "data analytics", "learning analytics", and "higher education" were considered. Figure 1. depicts the PRISMA flow diagram for the search and selection criteria for conducting the bibliometric analysis. An initial search was conducted using the keyword, abstract and publications, resulting in 613 articles/publications in the Scopus database. The number of articles was reduced to 481 after restricting the search period from 2015 to 2022 and including only articles written in English. The keywords were limited to include only the relevant keywords such as "machine learning", "learning analytics", "data analytics", "higher education", and "prediction". They excluded irrelevant keywords that were not relevant to the study. The 481 data were exported to an Excel CSV file for cleaning and further analysis. After assessing the titles and abstracts of the articles, 72 papers were identified as bibliometrically significant and therefore, bibliometric analysis was performed on them. The bibliometric data was analysed using Visualisation of Similarities (VOSviewer) and R-studio's biblioshiny.

The Visualisation of Similarities, also known as VOSviewer, is bibliometric visualisation software developed by Van Eck and Waltman in 2010. It can be used to construct and display bibliometric networks of authors, journals, papers, and countries/regions. The VOS software gathers data and generates maps based on bibliographic coupling, co-authorship, citation, co-citation, and keyword co-occurrence. VOSviewer employs two fundamental techniques: VOS mapping and VOS clustering. VOSviewer is available for download (available at http: //www. vosviewer.com).[28]

R-studio's biblioshiny is a web-based graphical interface developed by Massimo Aria and Corrado Cuccurullo. It works with WoS, Scopus and Dimensions. It includes analytics and graphs for three-level metrics (source, author and document) and three structures of knowledge which are conceptual, intellectual and social). The graphs and performance analysis generated can be exported to various file formats.[28]

Bibliometric Method

Bibliometric is the quantitative study of bibliometric content. The bibliometric method is classified into two main categories, namely: performance analysis and science mapping.[27] Performance analysis is mostly about discovering how important publications, authors and journals are in the scientific world based on the number of citations and publications. For example, the *h*-index or Journal Citation Reports (JCR) impact factors are considered in such analysis. Science mapping aims to identify a particular research field's "structural and dynamic" characteristics.[27] This is a way to determine how a research field is put together and how it changes over time. This is accomplished by building networks of elements based on the interconnectedness of the bibliometric materials and then categorising those elements into various groups and clusters. "Co-citation, "bibliometric coupling," and "co-occurrence of words" are the three primary techniques used in scientific mapping: "co-citation" studies a database's reference lists to determine the "knowledge base" or " intellectual structure".[29] The aim of "bibliometric coupling" is to establish a framework for future research, therefore the basis of how documents are cited and focused depends on the number of documents that share the same references, and lastly, the most frequently used keywords are analysed using words co-occurrence.

RESULTS

This section presents the findings from our bibliometric analysis. First, we present the characteristics of the dataset, followed by findings of the most influential authors and institutions. We continue by presenting the findings of the main themes and trends in the research area, and finally, the findings for collaboration patterns among authors and countries are presented.

Characteristics of Dataset

Table 1 below describes the main information in the Scopus database about predictive systems in education from 2017 to 2022. The information was generated using the R-studios biblioshiny software. Even though the search period specified was 2015 to 2022, the results indicate no publication from 2015 to 2016 in the Scopus database. The document sources were from only journals. The single-authored articles represent 15.78% of the total number of authors. The total number of co-authorships per article is 3.99%, whereas the total number of international co-authorships represents 36.11%. There were 246 author keywords, 13.71% average citation, 3408 references and an annual article growth rate of 78.26%.

From Table 1, it can be observed that single-authored articles were more than co-authored articles. The total number of single-authored per article is 6, whereas the total number of co-authored documents per article is 3.99. This indicates that there is less collaboration among researchers in the area of ML-based predictive systems. The number of single-authored articles was twice the number of co-authored articles. In addition, the total number of international co-authorship is lower at 36.11%, inferring that research collaboration between authors in different countries is lacking. However, the annual growth rate of 78.26 indicates an appreciable increase in the research area and authors' contribution over the years. Although the research annual growth rate appreciates, co-authorship among researchers is limited.

Influential authors and institutions

The result elucidates the most influential authors, most influential affiliations, annual scientific production, and total citations. In

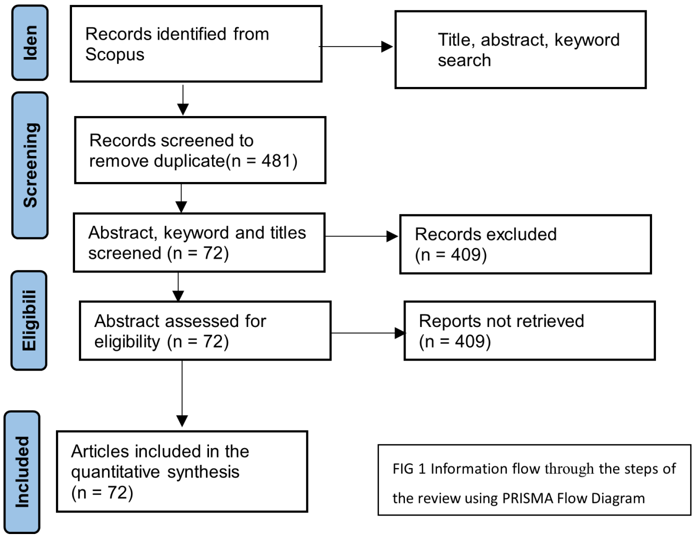
Iden
Records identified from Scopus
Title, abstract, keyword search
Screening
Records screened to remove duplicate(n = 481)
Abstract, keyword and titles screened (n = 72)
Records excluded (n = 409)
Eligibili
Abstract assessed for eligibility (n = 72)
Reports not retrieved (n = 409)
Included
Articles included in the quantitative synthesis (n = 72)
FIG 1 Information flow through the steps of the review using PRISMA Flow Diagram

Figure 1: Illustrate the search procedure to retrieve the relevant data from the Scopus database.[5]

order to accomplish this, Performance analysis was conducted to find out how important authors, affiliations and publications are in the research topic based on the number of citations and publications.

Figure 2 presents the 10 most relevant authors. Most relevant authors with publications on ML-based predictive systems in education include Moubayed A., Nassif AB, Shami A, and Zhang G. Four (4) out of the ten (10) most relevant authors have 3 publications whereas six (6) authors have two (2) publications each. Based on the findings presented in Table 2, it is evident that a total of 262 articles received the highest number of citations on a global scale. However, it is important to note that this does not necessarily indicate that the authors of those articles are the most relevant in their respective fields. The total number of the most cited Global articles was 262, and these were not necessarily the most relevant authors. However, the most cited authors happen to be authors who were not listed as the 10 most relevant authors. Ranking by the most cited author in Table 2, the top 10 researchers are Barlow C., Day C., Fan Z, Fuller A., Verbert K., Moubayed A., Nassif A.B, Shami A, Injadat M., Järvelä S. The 10 most relevant affiliations are presented in Figure 3. These include the Astec University of Southern Queensland, Nanjing University of Post and Telecommunication and the University of Ha'il, United Arab Emirates University, University of Sharjah, Arizona State University, Sharia University, Brigham Young University and Comsats University. The annual scientific production shown in Figure 4 stipulates the increase in research in predictive systems from 2019 to 2021, with 2021 representing the peak.

Performance Analysis

Number of publications of authors, citations and impact factors

In determining the most influential authors, the total link strength was considered among the most cited and the most relevant authors. The total link strength indicates the total strength of co-authorship links of a given researcher with other researchers. Ranking by the 10 most influential authors Moubayed A., Nassif A.B, Shami A., Injadat M., Deo R.C., Nguyen-Huy T., Yaseen Z.M., Verma C., Injadat M.N were presented. According to the findings, the authors with the most publications and co-authored links are the most influential in the research field.

Themes and trends in the research area

To further understand the ML-based predictive system in the higher education research field, we analysed the keywords and titles found in the dataset and the thematic evolution of the research area. The analysis on the co-occurrence of words was performed, focusing on the author's keywords and titles and thematic evolution. The author's keywords and title provide a better understanding of the research trends. The thematic evolution addresses the cluster and themes of the direction of the research field.

The following keywords in Figure 5 and trend analysis in Figures 6 and 7 generated using the R-studio biblioshiny provide insights into the topic and trends of the research area.

Figure 5 depicts a Wordcloud containing the most prevalent keywords. Data analytics appears to be the most popular keyword because it is presented in a bolder font than the others, followed by learning analytics, educational data mining, and others. Data analytics, learning analytics, and educational data mining are

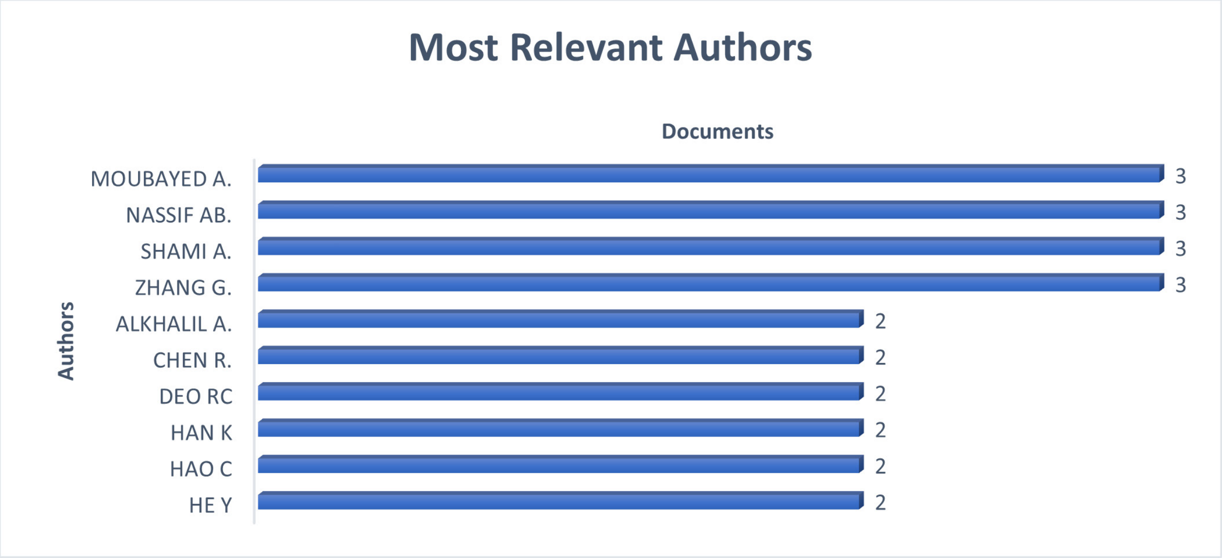

Figure 2: Most Relevant Authors.

the three most commonly used author keywords identified in this document. In learning analytics research, these terms are frequently used interchangeably (See.[30] Figure 6 and Figure 7 depict the most frequently used keywords and themes that reflect the current emphasis on predictive systems research in education. Based on document titles, the research themes include ML, data analytics, and learning analytics. These themes are further categorised under specific overarching themes, including the motor theme, the emerging theme, and the base theme. However, the motor themes reflect trends such as e-learning, artificial intelligence, the Gini-index, and the prediction of MOOCs. Emerging themes include deep learning and the Internet of Things, whereas the basic themes include ML data analytics, and learning analytics. Figure 7 depicts the current research trend as the "prediction." This demonstrates the intense research focus on prediction-related topics for 2021 and 2022.

However, it appears that there are fewer research terms in e-learning-related topics, as the most recently used terms span the years 2018 to 2020. In 2021, terms such as learning analytics, data analytics, and ML were prominent in most articles.

Collaboration patterns based on Authors and Countries

The countries and authors' collaboration analyses generated using the Vosviewer software provide an understanding of the research contributions by countries and the collaboration patterns of authors. To accomplish this, Science mapping was conducted to identify the number of citations, bibliometric coupling and co-citation analysis of authors and countries.

Figures 8 and 9 represent the countries cited most and co-authored in the research area. Ranking among the 10 most cited Countries is the United Kingdom, with a citation of 273, followed by Belgium with 155, and Canada with 72 citations. China had 60

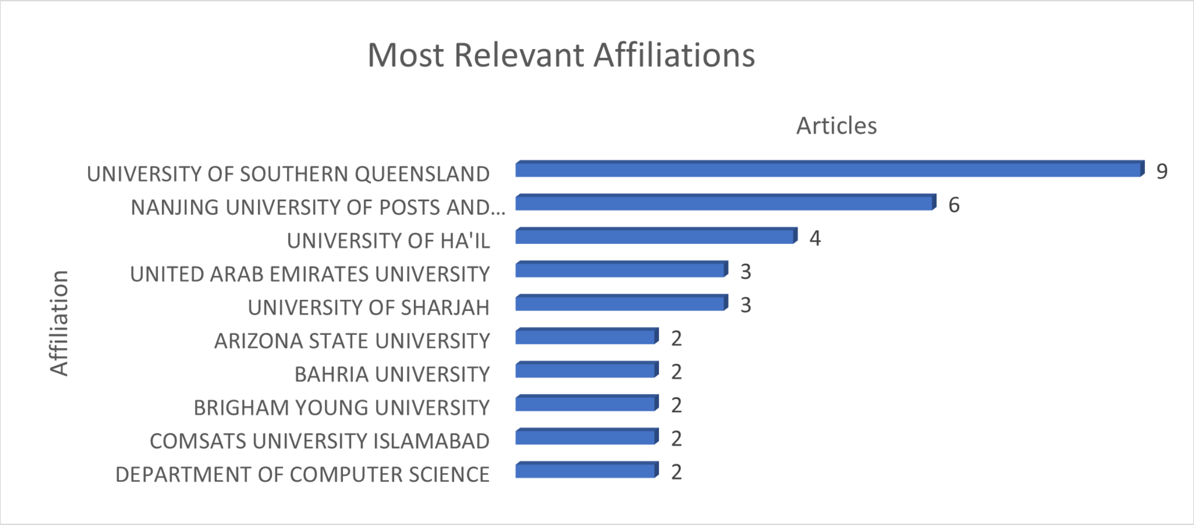

Figure 3: Most relevant Affiliations.

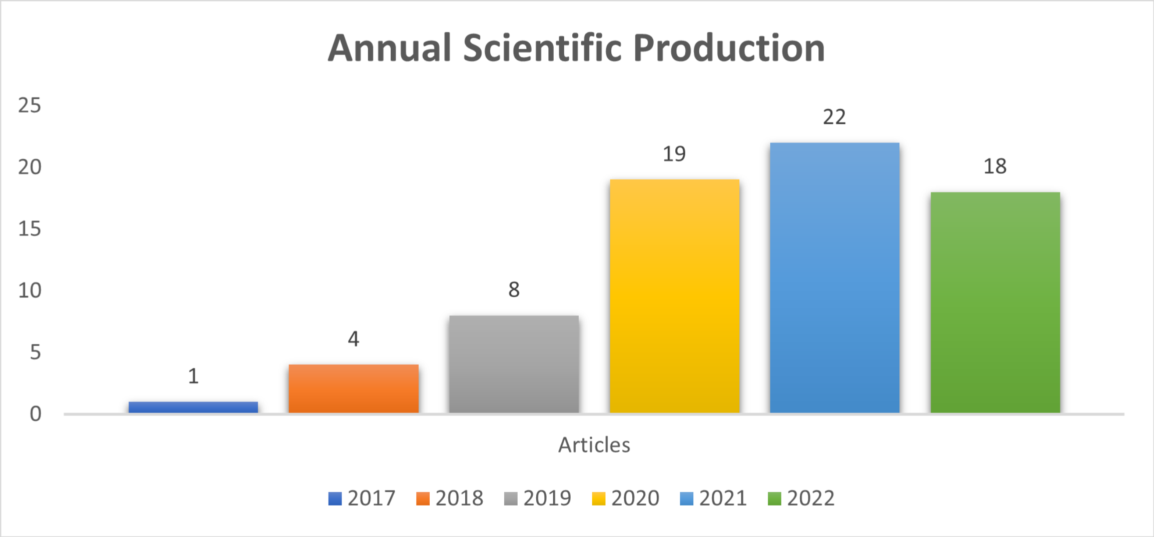

Figure 4: Scientific Impact of Authors.

publications, Korea with 50, Ecuador had 45, Hungary with 43, France with 37, followed by Ireland and Pakistan with 28 and 25 citations, respectively.

Figure 9 presents the bibliographic coupling of the countries' collaborations with other countries in the research area. The clusters indicate the concentrated countries with high total link strength. As presented in Figure 8, China, the United Kingdom, Pakistan, the United States of America, Spain, the United Arab and India are countries with high total link strength of coupling with other countries. From Figures 8 and 9 and Table 3, it can be deduced that the most cited countries are not necessarily the ones with more country collaborations; however, the higher the total link strength between countries, the more collaboration exists between them. Table 4 illustrate the collaboration between Australia, Bangladesh, Iraq, Malaysia and Sweden. The Next Country is Belgium, with collaborations between Ecuador, Finland, France and the Netherlands. Canada also collaborates with Gabon and Korea. Of all the collaborations, only one country is from Africa; thus, Gabon with an infinitesimal frequency. Even though research in predictive systems in education has gained popularity in developed countries, there is limited international collaboration among authors.

DISCUSSION

This bibliometric analysis aimed to identify the most relevant authors and affiliations in the domain of ML-based predictive systems. And also determine the research themes and trends to guide future research explorations and, finally, the collaboration among countries.

Regarding the most relevant authors and affiliation, the study's results indicate the significant growth in ML predictive systems research in education among countries such as the United

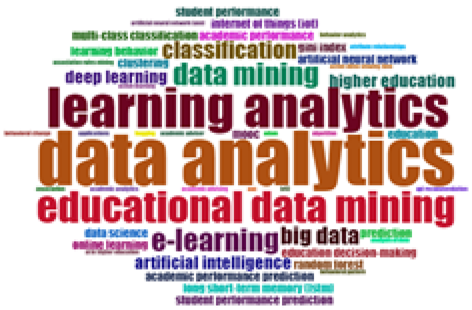

Figure 5: Wordcloud of author keywords.

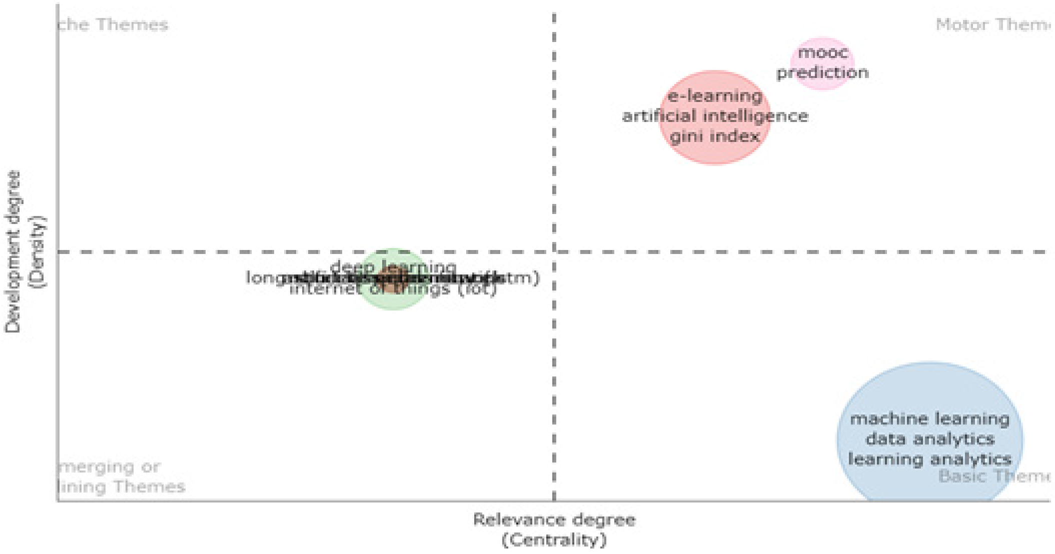

Figure 6: Thematic Evolution.

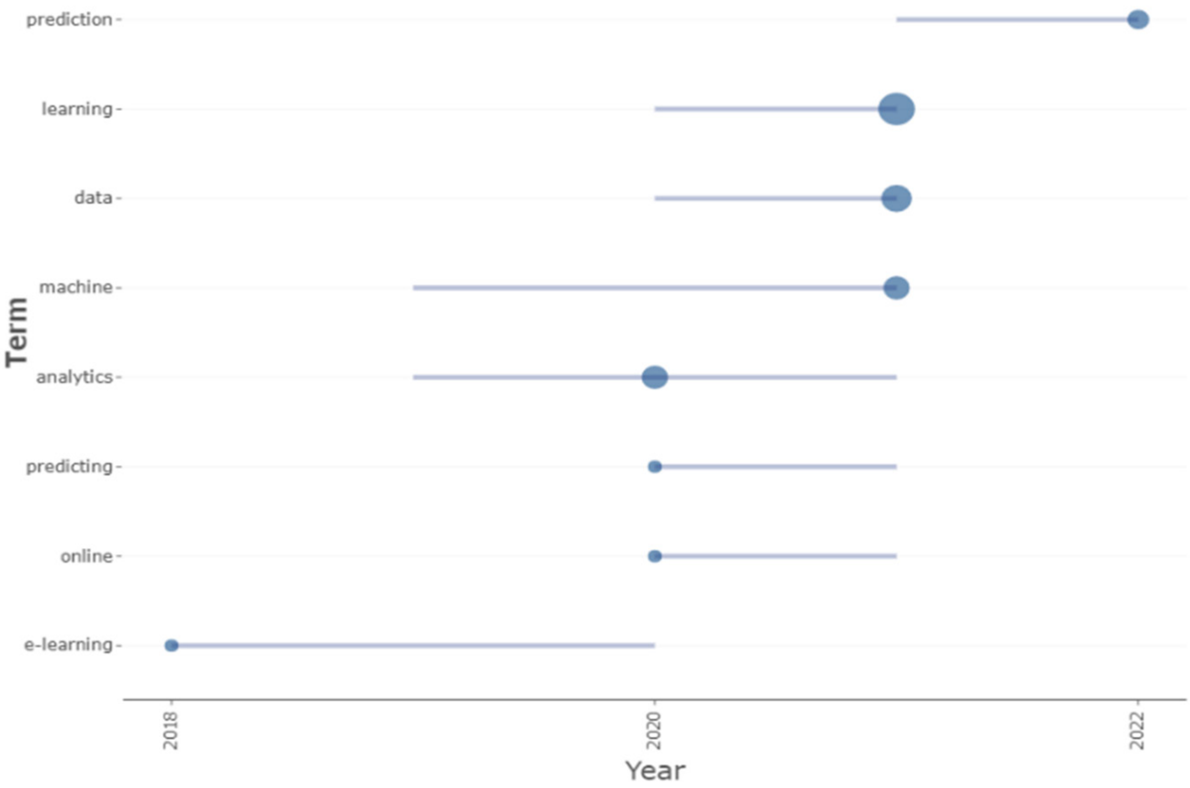

Figure 7: Trending Title.

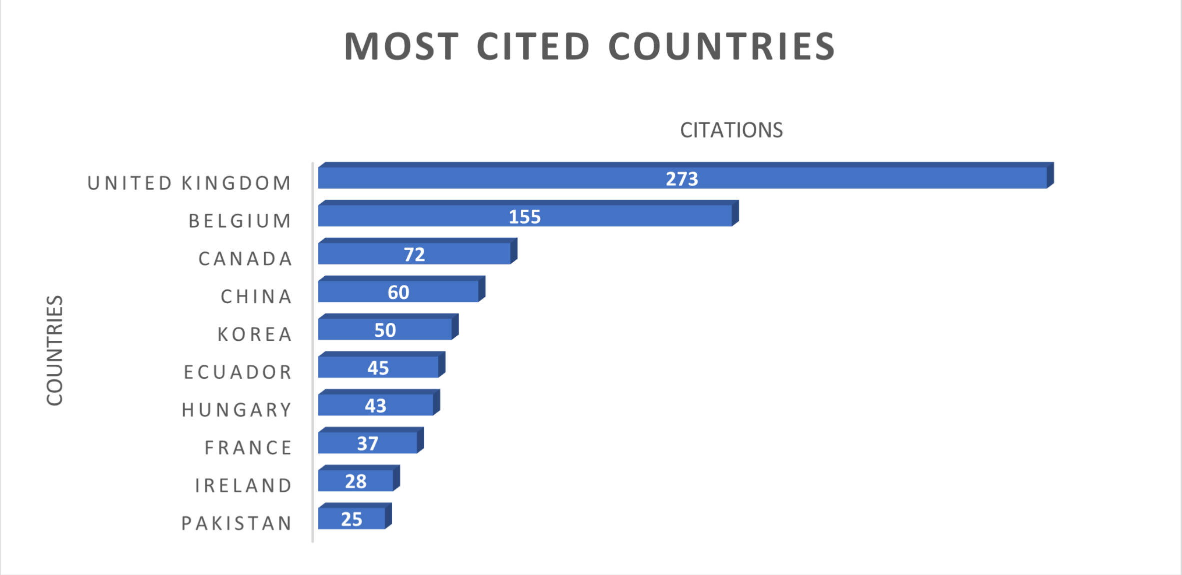

Figure 8: Most Cited Countries.

Kingdom, Belgium, Canada, China, Korea, Ecuador, Hungary, France, Ireland and Pakistan. The result also shows the most relevant authors as Moubayed A., Nassif AB, Shami A, and Zhang G. These authors are considered relevant because they have published articles in the field in the last two years, from 2020 to 2022. It is a fact that data-driven research in education has been ascending for the last two years, a phenomenon which can be attributed to the emergence of the internet and the high computing power available for data storage and analysis. In addition, many schools migrated to online teaching and learning during the COVID-19 pandemic and have continued to use such technologies to improve teaching and learning. As a result, the availability of digital footprint retrieved from users when accessing the internet enables data to be readily available.

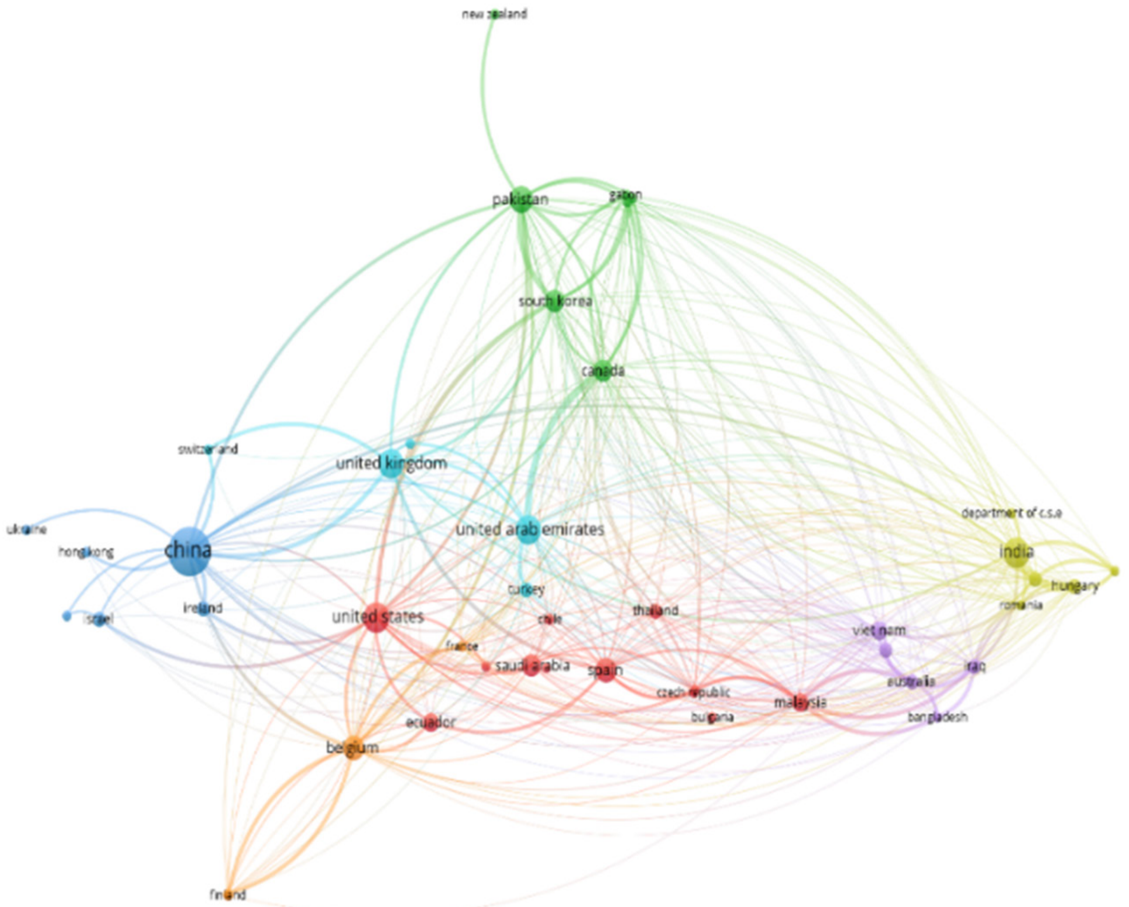

Figure 9: Bibliometric coupling of countries.

Therefore, performing analysis using readily available data is possible. However, research in ML-based predictive systems is widespread among authors and affiliations in developed countries such as the United Kingdom, Belgium and others. On the other hand, no affiliations or authors from Africa were featured in the 10 most relevant or cited authors.

Regarding themes and keywords, the results show that ML, Data Analysis and Learning Analytics were the most frequently used keywords among authors. The use of ML as the most used author keyword can be attributed to the emergence of Artificial Intelligence and the impact of research in ML. Moreover, keywords such as ML, Data Analysis, and Learning Analytics are mostly used interchangeably though there are slight distinctions due to the tools and focus of the methods applied (see[30]). On the research themes, the findings indicate clusters such as motor theme, emerging theme, base theme, and current trends. The research themes discovered at the motor theme include e-learning, artificial intelligence, Gini-index, and MOOC prediction. These themes are related as much research in ML predictive systems has made mention of e-learning systems or MOOC prediction using artificial intelligence tools. The emerging themes discovered in the findings show that research in predictive systems mainly focuses on deep learning and the Internet of Things (IoT). Imply The findings imply that more efficient Deep Learning algorithms would be applied in the design of predictive modelling to improve accuracy and performance. The emerging trends identified are similar to those identified in works[17,18] that researched smart education. It can be concluded that research trends in Artificial Intelligence and ML are advancing toward deep learning and IoT implementations. The base themes indicate the relevance of ML, Data Analytics, and Learning Analytics in the predictive system in education research. The findings also identified 'prediction' as the trending title in ML-based predictive system research. This implies that there is more research being undertaken in the area of predictive systems. However, the research concentration is confined to countries such as China United Kingdom, the United States and other developed countries. This implies a lack of research among African countries and authors in the study area.

Furthermore, the contribution of countries in the research field is an important factor as this determines the focus of the research domain based on countries. The most cited countries' results indicated that the United Kingdom is the most cited country in the research area. Also, the research collaboration among countries showed that Australia, Belgium, and Canada had co-authored publications with other countries on the topic, though it has less frequency. On the other hand, there is a lack

Table 1: Descriptive statistics

Main information about the data	Explanation	No.
Timespan	Year of publications	2017:2022
Sources (Journals)	Frequency distribution of sources	43
Documents	The number of articles	72
Annual Growth Rate %	Annual growth rate of articles in percentages	78.26
Document Average Age	Average age representation of articles	1.46
Average citations per doc	Average citation per article	13.71
References	Total number of references	3408
Author's Keywords (DE)	Total number of author keywords	246
Authors	Total number of authors	263
Authors of single-authored docs	Total number of authors of single-authored docs	6
Single-authored docs	Total number of authors per article	6
Co-Authors per Doc	Total number of co-authors per doc	3.99
International co-authorships %	Total number of international co-authorships %	36.11

Table 2: Most Cited Authors

Author	Documents	Citations
Barlow C.	1	262
Day C.	1	262
Fan Z.	1	262
Fuller a.	1	262
Verbert K.	2	116
Moubayed A.	3	109
Nassif A.B.	3	109
Shami A.	3	109
Injadat M.	2	89
Järvelä S.	1	73

Table 3: Bibliographic coupling of countries with total link strength

Country	Documents	Total link strength
China	20	312
United Kingdom	7	310
Pakistan	6	239
Belgium	5	202
United States	8	160
Spain	5	116
United Arab Emirates	7	108
India	8	91

of publications in the field of ML-based predictive systems in education in Africa. This could be attributed to Africa's lack of readily available educational data. Due to the partial use of LMS among some teachers, data about courses and assessments are stored on different systems, making it arduous to consolidate data for research. It was observed that research in Learning analytics is limited in Africa. Researchers and practitioners in Africa must dive into the research area in ML and learning analytics to generate

Table 4: Research collaboration between countries

From	To	Frequency
Australia	Bangladesh	1
Australia	Iraq	1
Australia	Malaysia	1
Australia	Sweden	1
Belgium	Ecuador	1
Belgium	Finland	1
Belgium	France	1
Belgium	Netherlands	1
Canada	Gabon	1
Canada	Korea	1

unique data about students and their learning environment to foster study success.

Research implication

First, this study provides a comprehensive review of the emerging topic of ML-based predictive systems in the higher education context. Secondly, the combination of Vosviewer and Bilioshiny tools for the analysis presents a methodological contribution to the field of bibliometric research. Thirdly, identifying the emerging research clusters provided by the keyword and thematic evolution analysis provides insight into the research direction in the study area. Lastly, the contribution of countries and collaboration analysis provided a knowledgeable insight into countries that fall short of research in this direction for which an actionable idea for research is proposed.

CONCLUSION AND RECOMMENDATION

The study analysed documents from the Scopus database using bibliometric analysis. It presented a descriptive analysis of the total number of articles retrieved, the total number of authors, the total number of references, single-authored publications, and co-authored information. The search string included the period from 2015 to 2022, and the results of the analysis presented the period from 2017 to 2022, which translates as a shortfall in research and publication in ML-based predictive systems during the years 2015 and 2016 in the Scopus database. The study identified the most influential authors, affiliations, and countries in the research area by analysing the most relevant authors, most relevant affiliations, and authors' production growth. The keyword and trend analysis were conducted to identify the most common authors' keywords. The keywords most frequently used were ML, Data Mining, and Learning Analytics. The themes and trends were classified into the motor, emerging, and basic themes. While the motor theme showed predictive systems as the trending topic, the emerging trends revealed that deep learning and IoT are the focus of research in the future of Predictive systems in education. Furthermore, the study discovered a lack of research in the subject area in Africa. There was limited collaboration among countries except for Australia, Belgium, and Canada, which have research collaboration among other countries. However, only one research collaboration on the research focus was identified in Africa.

This study serves as a basis for researchers and educational stakeholders in Africa to focus on research into predictive systems using ML and Internet of Things (IoT) technology to develop more comprehensive models to improve student performance in higher educational institutions.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Sandra L, Lumbangaol F, Matsuo T. Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review. TEM Journal. 2021;10(4):1919-27.
2. Tahiru F. AI in education: A systematic literature review. Journal of Cases on Information Technology. 2021;23(1):1-20.
3. Tahiru F, Agbesi S. The Future of Artificial Intelligence in Education. In: Digital Technology Advancements in Knowledge Management. IGI Global.; 2021. p 187-94.
4. Chung JY, Lee S. Dropout early warning systems for high school students using machine learning. Child. Youth Serv. Rev. 2019;96:346-53.
5. De Oliveira TN, Bernardini F, Viterbo J. An Overview on the Use of Educational Data Mining for Constructing Recommendation Systems to Mitigate Retention in Higher Education. In: IEEE Frontiers in Education Conference. 2021:1-7.
6. Umer R, Susnjak T, Mathrani A, Suriadi L. Current stance on predictive analytics in higher education: opportunities, challenges and future directions. Interact. Learn. Environ. 2021.
7. Ahsan M, Siddique Z. International Journal of Information Management Data Insights Industry 4. 0 in Healthcare : A systematic review. Int J Inf Manag Data Insights [Internet] 2022;2(1):100079. Available from: https://doi.org/10.1016/j.jjimei.2022.100079
8. Yang Y, Chen G, Reniers G, Goerlandt F. A bibliometric analysis of process safety research in China:Understanding safety research progress as a basis for making China's chemical industry more sustainable. J Clean Prod. 2020.
9. dos Santos BS, Steiner MT, Fenerich AT, Lima RH. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. Computers & Industrial Engineering. 2019;138:106120.
10. Pears M, Konstantinidis S. Bibliometric Analysis of Chatbots in Health-Trend Shifts and Advancements in Artificial Intelligence for Personalised Conversational Agents. 2022;2020:494-8.
11. Firdaus A, Faizal M, Feizollah A, Abaker I, Hashem T. The rise of "blockchain": bibliometric analysis of blockchain study. Scientometrics. 2019;120:1289-331.
12. Xu X, Chen X, Jia F, Brown S, Gong Y, Xu Y. Supply chain finance: A systematic literature review and bibliometric analysis. International Journal of Production Economics. 2018;204:160-73.

13. Djeki E, Dégila J, Bondiombouy C, Alhassan MH. E-learning bibliometric analysis from 2015 to 2020. Journal of Computers in Education. 2022;9(4):727-54. DOI: 10.1007/s40692-021-00218-4
14. Shukla AK, Muhuri PK, Abraham A. Engineering Applications of Artificial Intelligence A bibliometric analysis and cutting-edge overview on fuzzy techniques in Big Data. Engineering Applications of Artificial Intelligence. 2020;92:103625.
15. Zhao L, Dai T, Qiao Z, Sun P, Hao J, Yang Y. Application of artificial intelligence to wastewater treatment : A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. Process Safety and Environmental Protection. 2020;133(92):169-82.
16. Ali M, Rahimi R, Okumus F, Liu J. Bibliometric studies in tourism. Annals of Tourism Research. 2016;61:180-98.
17. Guo XR, Li X, Guo YM. Mapping knowledge domain analysis in smart education research. Sustainability. 2021;13(23):13234.
18. Li KC, Wong BT. Research landscape of smart education: a bibliometric analysis. Interactive Technology and Smart Education. 2022;19(1):3-19.
19. Suhaimi NA, Kamaruddin NB, Subramaniam TT, Sjarif NN, Masrom MB, Maarop NB. Classification of Learner Retention using Machine Learning Approaches. In: 7th International Conference on Research and Innovation in Information Systems (ICRIIS 2021). 2021;0-5.
20. Prudencio RF, Maximo MR, Colombini EL. A survey on offline reinforcement learning: Taxonomy, review, and open problems. IEEE Transactions on Neural Networks and Learning Systems. 2023. Available from: http://arxiv.org/abs/2203.01387
21. Yousafzai BK, Hayat M, Afzal S. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. Education and Information Technologies. 2020;25(6):4677-97.
22. Sassirekha MS, Vijayalakshmi S. Predicting the academic progression in student's standpoint using machine learning. Automatika. 2022;63(4):605-17.
23. Liu Y. Evaluation Algorithm of Teaching Work Quality in Colleges and Universities Based on Deep Denoising Autoencoder Network. Mob Inf Syst. 2021;2021.
24. Kehinde AJ, Adeniyi AE, Ogundokun RO, Gupta H, Misra S. Prediction of Students' Performance with Artificial Neural Network Using Demographic Traits. Lect Notes Electr Eng. 2022;855:613-24.
25. Liu X, Ardakani SP. A machine learning enabled affective E-learning system model. Education and Information Technologies. 2022;
26. Oralbayeva N, Shakerimov A, Sarmonov S, Kantoreyeva K, Dadebayeva F, Serkali N, *et al*. K-Qbot : Language Learning Chatbot based on Reinforcement Learning. 2017.
27. Piñeiro-Chousa J, López-Cabarcos MÁ, Romero-Castro NM, Pérez-Pico AM. Innovation, entrepreneurship and knowledge in the business scientific field: Mapping the research front. Journal of Business Research. 2020;115:475-85.
28. Liao H, Tang M, Li Z, Lev B. Bibliometric analysis for highly cited papers in operations research and management science from 2008 to 2017 based on essential science indicators. Omega. 2019;88:223-36.
29. Mariani M, Borghi M. Industry 4.0: A bibliometric review of its managerial intellectual structure and potential evolution in the service industries. Technological Forecasting and Social Change. 2019;149:119752.
30. Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery. 2020;10(3):e1355.

Cite this article: Tahiru F, Parbanath S, Agbesi S. Machine Learning-based Predictive Systems in Higher Education: A Bibliometric Analysis. J Scientometric Res. 2023;12(2):436-47.