A critical review of data mining in education on the levels and aspects of education

Inusah Fuseini
Department of Mathematics and ICT Education, University for Development Studies, Tamale, Ghana and Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, and

Yaw Marfo Missah
Department of Computer Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

Abstract

Purpose – This systematic literature review aims to identify the pattern of data mining (DM) research by looking at the levels and aspects of education.

Design/methodology/approach – This paper reviews 113 conference and research papers from well-known publishers of educational data mining (EDM) and learning analytics-related research using a recognized literature review in computer science by Carrera-Rivera et al. (2022a). Two major stages, planning and conducting the review, were used. The databases of Elsevier, Springer, IEEE, SAI, Hindawi, MDPI, Wiley, Emerald and Sage were searched to retrieve EDM papers from the period 2017 to 2023. The papers retrieved were then filtered based on the application of DM to the three educational levels – basic, pre-tertiary and tertiary education.

Findings – EDM is concentrated on higher education. Basic education is not given the needed attention in EDM. This does not enhance inclusivity and equity. Learner performance is given much attention. Resource availability and teaching and learning are not given the needed attention.

Research limitations/implications – This review is limited to only EDM. Literature from the year 2017 to 2023 is covered. Other aspects of DM and other relevant literature published in EDM outside the research period are not considered.

Practical implications – As the current trend of EDM shows an increase in zeal, future research in EDM should concentrate on the lower levels of education to identify the challenges of basic education which serves as the core of education. This will enable addressing the challenges of education at an early stage and facilitate getting a quality education at all levels of education. Appropriate EDM techniques for mining the data at this level should be the focus of the research. Specifically, techniques that can cater for the variation in learner abilities and the appropriate identification of learner needs should be considered.

Social implications – Content sequencing is necessary in facilitating an easy understanding of concepts. Curriculum design from basic to higher education dwells much on this. Identifying the challenge of learning at
the early stages will facilitate efficient learning. At the basic level of learning, data on learning should be collected by educational institutions just as it is done at the tertiary level. This will enable EDM to accurately identify the challenges and appropriate solutions to educational problems. Resource availability is a catalyst for effective teaching and learning. The attributes of a learner will enable knowing the true nature of the learner to determine the prospects of the learner.

Originality/value – This research has not been published in any journal. The information presented is the original knowledge of the authors. However, a pre-print of the work is in Research Square.

Keywords Basic education, Data mining, Tertiary, Pre-tertiary, Teaching and learning

Paper type Literature review

1. Introduction
 Interestingly, data mining (DM) in education has played a tremendous role in the application of technology in education helping educational leaders to make accurate decisions (Inusah et al., 2023a). To enhance more accuracy and consistency in educational data for efficient management, DM is needed (Inusah et al., 2023b). Specifically, educational data mining (EDM) is the process of revealing hidden information from larger educational data for effective decision-making, while learning analytics looks at pedagogical content and the processes that yield to an outcome in learning. Learning analytics emphasises learning and the outcome (Lemay et al., 2021). The significance of EDM in education is well noted in academic performance of learners through resources management, learner attributes. The application of EDM is therefore a necessity for life-long learning. As a novel methodology for solving problems in education, much research conducted in this area is basically on pedagogy which enriches the quality of tuition (Sokkhey et al., 2020). Deep learning as a trending EDM technique has gained much attention in tracing patterns and detailed solutions to educational problems (Hernández-Blanco et al., 2019). In a three-decade analysis of EDM, the application of clustering algorithms and their applications, as well as the usability of EDM, has been reviewed (Ashraf et al., 2020). Enhancing more accuracy of using this methodology is necessary. Research by Tsiakmaki et al. (2020) considers determining accuracy using the rightful DM techniques. Gamification as a trend for arousing the interest of learners is a novel technique that is combined with EDM to spice up effective teaching and learning (Daghestani et al., 2020; Arend et al., 2022). In Issah et al. (2023), the attributes determining the performance of learners using DM techniques are reviewed.

The purpose of this review is to find out the levels of education given attention in EDM and the aspect of education which is usually considered. The research looks at three educational levels: basic education, secondary education and tertiary education. Basic education which is the lower level of education looks at the first 14 years of education. It is seen as the foundation (core) of education. Secondary education looks at the institutions after basic education which is non-tertiary. Tertiary education denotes the higher levels. The various aspects of education that are usually considered at the various levels consist of teaching and learning activities as well as the resources for learning which yields the performance in education. Attributes of learners and many other aspects of learning are considered. Overwhelming research on EDM stresses the attributes of the learner in determining academic performance. The following research questions guide the review:

**RQ1.** What is the current trend in EDM that affects effective teaching and learning?

**RQ2.** Which level of education is given much attention in EDM research?

**RQ3.** Which aspect of education is given much attention in EDM to enhance effective learning?
These questions are necessary because the focus of recent trends in EDM regarding levels and aspects of education is the foundation for higher education. Also, resource availability and learner attributes contribute significantly to a learner’s academic growth and development. The remaining part of the review is organized into six areas. The first part is the related reviews on EDM which comprises the EDM in education, the levels of education (basic, secondary and tertiary) and the major aspects of education (academic performance, teaching and learning resources and learner attributes). The second aspect is the methodology used in the literature search criteria which comprises how the search was conducted and the search strategy. The publication selection looks at inclusion and exclusion criteria. The third aspect is the presentation of the results for the literature used. The fourth aspect is the discussion of the research gap identified which looks at both the population gap and the knowledge gap. The fifth aspect is the conclusions drawn and the sixth is the recommendations to stakeholders of education. Table 1 is related to reviews in EDM.

2. Related reviews on educational data mining
In Ashraf et al. (2020) and Dutt et al. (2017), the use of clustering algorithms and their application in EDM is specified using a trend of three decades. The trends of DM shortly are also included in this work. Deep learning approaches to EDM are reviewed to know the task that has already been taken and the future directions (Hernández-Blanco et al., 2019). At the higher education level, a comprehensive review of EDM is conducted to enable potential authors to know the effects on students’ performance and the possible direction to improve effective teaching and learning (Aldowah et al., 2019). Focusing on only classroom learning, (Khan and Ghosh, 2021) used necessary predictors of time, methods and aims to identify significant performance predictions during a course. In a related paper, Alshareef et al. (2020) recommend using extended tools to measure and predict students’ performance in teaching and learning. Using classification models, two attributes – grades and absence – are seen as the most significant in determining the performance of students in a year (Fernandes et al., 2019). Video analytics and DM techniques are used in predicting students’ performance at the higher education level to know the level of accuracy of the techniques applied (Hasan et al., 2020). In Du et al. (2020), extensive review of EDM at the various levels of education reveals the rapid application of DM techniques across the levels of education but rare information available on K–12 which is the basic or early level of education. The potential influence of DM in identifying the learning process and progress is identified in higher education (Aldowah et al., 2019).

3. Methodology
This paper reviews 113 conference and research papers from well-known publishers of EDM and learning analytics-related research using a recognized literature review in computer science by Carrera-Rivera et al. (2022). Two major stages, planning and conducting the review, were used. The databases of Elsevier, Springer, IEEE, SAI, Hindawi, MDPI, Wiley, Emerald and Sage were searched to retrieve EDM papers from the period 2017 to 2023. The papers retrieved were then filtered based on the application of DM to the three educational levels; Basic, pre-tertiary and tertiary education. Furthermore, the papers were filtered based on three aspects of education; student performance, teaching and learning resources and attributes of learners. In all, 20 conference papers and 93 research articles met the criteria.
<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Years covered</th>
<th>Databases explored</th>
<th>Papers used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutt et al. (2017)</td>
<td>A systematic review on educational data mining</td>
<td>1983–2016</td>
<td>IEEE Explore, ACM Digital, JEDM (Journal of Educational Data Mining), ProQuest and ScienceDirect</td>
<td>166</td>
</tr>
<tr>
<td>Chango et al. (2022)</td>
<td>A review of data fusion in multimodal learning analytics and educational data mining</td>
<td>Earlier–2021</td>
<td>Google Scholar, Web of Science and Scopus</td>
<td>42</td>
</tr>
<tr>
<td>Okewu et al. (2021)</td>
<td>Artificial neural networks for educational data mining in higher education: a systematic literature review</td>
<td>2010–2018</td>
<td>Web of Science, Scopus, IEEE Xplore, ACM Digital Library and Springer link</td>
<td>190</td>
</tr>
</tbody>
</table>

Source: Authors’ construct
3.1 Planning
After the initial idea formulation, six major steps were considered in the planning of the literature search. The first step was to define the PICOC and synonyms for clarity and accuracy in identifying the literature. The second step was to formulate the research questions that guide the literature review. This was followed by the selection of digital library sources. The fourth step was to define the inclusion and exclusion criteria. The fifth stage was to design the quality assessment checklist and the last step was designing the data extraction form.

3.2 Conducting the search
After the planning of the literature stage, the next stage is the conducting of a literature search. This comprises five major steps. The first step is to build digital library search strings. This was to get the specific publications from the databases selected. The second step was to gather the studies from the digital libraries using the search strings specified in the first step. All the nine databases selected were thoroughly searched to retrieve the information. The third step was to do a study selection and refinement. Three conditions were used in this stage for selection. The first condition was to ensure there was no duplicate. The next was to evaluate the inclusion and exclusion criteria and the last was to assign quality scores. After these conditions were met in the third step, the fourth step was to do data extraction from the selected papers and a final analysis and report presentation was done in the last stage.

To identify relevant papers for the criteria, a search phrase was necessary. The phrase “educational data mining” or “data mining in education” was used to locate the related literature. Further filtering by specifying the years of publication was done to limit the retrieved papers to the most relevant works from the year 2017 to 2023. Papers that were applied to the various levels of education were further filtered by manually reading the abstract and the methodology to know what the paper was about. Concentration was given to the aspect of education applied in the research; student performance, teaching and learning resources and learner attributes.

3.2.1 Search strategies. A Boolean search was used to locate all necessary publications before filtering. The search string included OR, NOT and AND to enable specifically getting the relevant literature. In all nine databases of well-known publishers, a thorough search was conducted. The nine publishers and the literature rating are in Table 1. The search strings used in the search can be seen below (“educational data mining” or “data mining in education” or “education and data mining”) AND (“learner attributes” or “performance” or “level of learning”) NOT (“learning systems”).

3.3 Publication selection
Duplicates were checked to ensure the same information from different databases was not included more than once. This was done by the reading of titles, abstracts and the authors as well as the year of publication.

3.3.1 The inclusion criteria. The most relevant papers to the research were from only conferences and articles from journal publications in the specified publishers. Specifically, research on how EDM is applied to a level of education and an aspect of education was the inclusion criteria.

3.3.2 Exclusion criteria. Three significant steps were used to exclude research works that were not most recently published and relevant to the research.

- Papers that are not relevant to the objectives of the research.
3.3.3 Selecting primary sources. The selection process included the string “educational data mining” and the use of keywords; education, data and mining. Papers that did not have the string phrase but included the keywords specified were thoroughly read to know the full-text content and include or exclude. To avoid errors and duplications, the primary reviewer’s work was reviewed by a second reviewer to ascertain its accuracy and precision. Random selection of the papers under the various categorizations was used as a criterion by the second reviewer. This methodology is the adoption of Carrera-Rivera et al. (2022) guide for systematic literature review in computer science. Fig 1 is the flowchart.

4. Results
Revelations from the investigation into the already existing literature produced an interesting trend for publications in EDM. For clarity and consistency in the result presentation, the results and findings of the review are presented in line with the research questions. Table 2 is the source and results of the literature search.

![Flowchart](image-url)

**Figure 1.** Flowchart for literature guide in a systematic review

**Source:** (Carrera-Rivera et al. 2022)
From Table 2, the number of databases searched for the literature totalled nine. These databases are for identified academic publishers of EDM for this work. The total number of responses gotten from the search is 536. After the first filtering, 317 papers were qualified. When the final filtering was done, only 113 papers were qualified to be used in the review. These papers were those that met the inclusion criteria specified in the methodology.

Table 2 is the categorization of the qualified papers into the levels and aspects of education. Both the basic level which is seen as the core has insignificant publications of nine which is about 7.96%, the pre-tertiary (secondary) also has nine which is about 7.96% and the tertiary has 95 which is about 84%. Specifically, for the aspects of education, academic performance at the basic level recorded only one publication, teaching and learning resource management recorded seven and learner attributes recorded only one. For the pre-tertiary level, academic performance was recorded six, teaching and learning resource management recorded one and learner attributes recorded two. The tertiary level which received overwhelming recognition had 47 for academic performance, 32 for teaching and learning resource management and 16 for learner attributes.

5. Discussion

Conscious efforts are made by researchers such as Inusah et al. (2021) in tracing problems of education using EDM as indicated in Table 3. However, the current trend in the rise of EDM with the relegation of basic education raises a concern (see Figures 2 and 3). This supports the Jiang (2022) assertion that researchers do not have enough information to predict the challenges of basic education. With Yang et al. (2022) alone spotted as the literature for performance and Francis and Babu (2019) as the only literature on learner attributes at the basic level, the results presented in Table 3 confirm the insufficient publications though the publications are limited to full manuscript journal research and conference papers. There is enough evidence to believe that the basic level is relegated. Resource management and utilization are essential for effective teaching and learning (Inusah et al., 2022). The basic (foundation) of the success of education in every nation has an insignificant number of research papers in EDM. Challenges of education can be identified at this level of education for lifelong learning of the citizenry. Accurate planning and prediction of the needs of learners in the future can be done to improve teaching and learning. If the foundation is weak, then the entire structure of education will be weaker. Learner attributes must be carefully identified at this stage and used to improve learning for good education. Proper storage of information in schools at the higher levels of education in databases makes it
<table>
<thead>
<tr>
<th>Level</th>
<th>Aspect</th>
<th>Published papers</th>
<th>No. of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>Academic performance</td>
<td>Yang et al. (2022)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Teaching and learning resource management</td>
<td>Mushtaq et al. (2019); Inusah et al. (2022); Inusah et al. (2021); Jiang (2022); Tsiakmakí et al. (2020); Torcate and De Oliveira Rodrigues (2021); Marfo Missah et al. (2023)</td>
<td>7</td>
</tr>
<tr>
<td>Pre-tertiary</td>
<td>Learner attributes</td>
<td>Francis and Babu (2019)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Academic Performance</td>
<td>Amjad et al. (2022); Abu Zohair (2019); Rebai et al. (2020); Livieris et al. (2018); Abideen et al. (2023); Alghamdi and Rahman (2023)</td>
<td>6</td>
</tr>
<tr>
<td>Tertiary</td>
<td>Academic performance</td>
<td>Francis and Babu (2019); Asif et al. (2017); Haryani et al. (2022); Yağca (2022); Dabhade et al. (2021); Hung et al. (2020); Aldowah et al. (2019); Kurdi et al. (2018); Trakunphutthirak and Lee (2021); Hassan et al. (2020); Wu (2022a, 2022b, 2022c); Hassam et al. (2024); Feng et al. (2022); Abu Zohair (2019); Sokkhey et al. (2020); Tsiakmakí et al. (2020); Fernandes et al. (2019); Tasnim et al. (2020); Al-Barrak (2016); Adejo and Connolly (2018), Adekitan et al. (2019); Cardona et al. (2023); Ashraf et al. (2020); Chaturvedi (2017); Roslan et al. (2023); Ordóñez-Avila et al. (2023); Sánchez et al. (2023); Chytas et al. (2023); Yurum et al. (2023); Brandon (2023); Tan (2023); Sun (2023); Sghir et al. (2023); Guleria and Sood (2019); Adekitan and Noma-Osaghae (2019); Lin et al. (2023); Rodrigues et al. (2018); Khan and Ghosh (2021); Hung et al. (2020); Arfaee et al. (2022) (Chu et al. 2022); Wu (2022a, 2022b, 2022c); Li et al. (2022); Zhao (2022); Wu (2022a, 2022b, 2022c); Peng and Tang (2022); Shafiq et al. (2022); Zhou and Wu (2022); Prada et al. (2020); Daoudi et al. (2021); Mazidi and Abusham (2018); Tang et al. (2018); Chen (2019); Du et al. (2023); H. Wang (2023); Du et al. (2022); Tang et al. (2018); Biao (2023); Contreras-Luján et al. (2022); Zhang and Qin (2018); Jiang and Zeng (2019); Lv (2021) (Hu and Guo (2018); Shayakhmetova et al. (2021); Yang (2021); Bowen (2021); Paredes et al. (2021); Joshy et al. (2023); Shaik et al. (2023); Alam (2023); Nascimento et al. (2018); Leitner et al. (2019); Martínez-Abad et al. (2020); Xing and Du (2019); Cardona et al. (2020); Hung et al. (2020); Wang et al. (2020); Davies et al. (2021); Xilin and Honglian, 2021); Wang, 2018; Aldowah et al. (2019); Mohd Tahib et al. (2023); Rafique et al. (2023); Zárate et al. (2023); Gera et al. (2024); Wongvorachan et al. (2023)</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Tea§ing and learning resource management</td>
<td>Shao et al. (2020), Al-Barrak (2016)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Learner attributes</td>
<td>Nascimento et al. (2018); Leitner et al. (2019); Martínez-Abad et al. (2020); Xing and Du (2019); Cardona et al. (2020); Hung et al. (2020); Wang et al. (2020); Davies et al. (2021); Xilin and Honglian, 2021); Wang, 2018; Aldowah et al. (2019); Mohd Tahib et al. (2023); Rafique et al. (2023); Zárate et al. (2023); Gera et al. (2024); Wongvorachan et al. (2023)</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 3. EDM as applied in levels and aspects

Source: Authors’ construct
relatively easier to analyse data. Pre-tertiary level of education has meaningful data available from learners and this information can be used to predict the likely occurrence of events in the schools using the attributes of the learner (Shao et al., 2020) As spotted by Amjad et al. (2022), Yağcı (2022) and Khan and Ghosh (2021) among others, the academic
performance of learners at this level of education can easily be predicated using other factors. With only two spotted literature on resource management (Arfaee et al., 2022; Rahman et al., 2022a, 2022b), EDM for effective management of resources is insufficiently utilized. Unfortunately, this level of education is not given much attention in EDM. As the progress of learners to tertiary institutions is best determined at this level, the use of EDM to identify the appropriate programme a learner can pursue at the tertiary level should be done with ease. Challenges of education at this level can help both the basic and tertiary levels. Learners’ performance at this level can help stakeholders in basic education to know how to prepare learners while the universities will also know how to deal with these learners.

Tertiary institutions are well organized and properly structured to make predictions easier (Francis and Babu, 2019; Yağcı, 2022) which makes them suitable for EDM. This could be the reason for the overwhelming research in EDM at this level. A larger chunk of research on EDM is on tertiary institutions where independent learners and their attributes are used in predicting academic performance. In all aspects of EDM, the tertiary level has a majority of publications. This could also be attributed to the universal nature of the tertiary system. However, the research works at this level failed to look at the previous performance and attributes of learners at the secondary level and its effects on the learner at the tertiary level. Also, revisiting learners who complete institutions and return for further studies should be identified. This will help universities to know the true needs of the learner and the demands of society. As resources are very necessary for effective delivery, conscious efforts by researchers to predict the effects of resource availability on the performance of the learner are necessary (Inusah et al., 2022). This will help institutions to carefully plan and provide educational resources that can boost the performance of learners. The attributes of learners which also affect the performance of learners should be looked at in the direction of programme choice which affects the occupations for employment. This will enable us to know the attributes associated with a particular occupation. In totality, the overwhelming research on EDM centred on the academic performance of learners which in some cases are replication and repetition of research works should be given a different direction. The factors determining the performance of a learner should be given a chance in EDM. The population gap identified in this research is the learners at the basic level. These groups of learners are neglected in EDM both in attributes and academic performance. They are not also considered in the provision and utilization of teaching and learning resources for effective learning. This creates a bigger challenge in education as they are not adequately prepared for the pre-tertiary and tertiary levels. The knowledge gap in this research is the limitation in the application of DM in education. Concentration is on the academic performance of learners with little or no attention to the learner attributes and teaching and learning resource management. Also the techniques used are limited to academic achievements with little or no information on the possibility of dropout or the cost of educating the learner.

5.1 Theoretical contribution
Content sequencing is necessary to facilitate an easy understanding of concepts. Curriculum design from basic to higher education dwells much on this. Identifying the challenge of learning at the early stages will facilitate efficient learning. At the basic level of learning, data on learning should be collected by educational institutions just as it is done at the tertiary level. This will enable EDM to accurately identify the challenges and appropriate solutions to educational problems. Resource availability is a catalyst for effective teaching and learning. The attributes of a learner will enable knowing the true nature of the learner to determine the prospects of the learner.
5.2 Practical implication
As the current trend of EDM shows an increase in zeal, future research in EDM should concentrate on the lower levels of education to identify the challenges of basic education which serves as the core of education. This will enable addressing the challenges of education at an early stage and facilitate getting a quality education at all levels of education. Appropriate EDM techniques for mining the data at this level should be the focus of the research. Specifically, techniques that can cater for the variation in learner abilities and the appropriate identification of learner needs should be considered.

6. Conclusion
This paper presents a thorough review of EDM by considering conference papers and articles from nine well-known publishers. All levels of education and research work are considered to know the application DM and the aspect of the application. The pattern of EDM is identified and the research gaps that need more attention in the area of EDM are identified. In summary, the tertiary level of education is given much attention in EDM at the expense of basic education for growth and development in education. The academic performance of learners is the most recognized aspect of education by researchers with little recognition of resources for a learner. Careful application of EDM at the lower levels of education can help in identifying the challenges of education to gain sustainable solutions for the improvement of the education sector and the development of nations as a whole.

7. Recommendation
Stakeholders in education must recognize basic education as the pivot for education and assist in the utilization of EDM methodologies to accurately identify the challenges and the appropriate interventions at this level. This is necessary to help enhance efficiency in educational management for effective teaching and learning using the available limited resources. It will also help in leveraging the success of education to other levels of education. Adequate provision of resources at this level will help in laying a foundation for the future of education. It will enhance inclusivity and equity to enable every child of school-going age to gain quality education. The progression of children of school-going ages will also be smooth as they have a good foundation of learning.

References


Further reading

Corresponding author
Inusah Fuseini can be contacted at: finusah@uds.edu

Data mining in education

59

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com